The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/0265-671X.htm

International Journal of Quality & Reliability Management

QUALITY PAPER

An integrated data-driven framework for vehicle quality analysis based on maintenance record mining and Bayesian network

Received 30 March 2023
Revised 14 June 2023
Accepted 23 April 2024

Aoxiang Cheng and Youyi Bi
University of Michigan – Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai, China

Abstract

Purpose – The purpose of this paper is to present an integrated data-driven framework for processing and analyzing large-scale vehicle maintenance records to get more comprehensive understanding on vehicle quality.

Design/methodology/approach – We propose a framework for vehicle quality analysis based on maintenance record mining and Bayesian Network. It includes the development of a comprehensive dictionary for efficient classification of maintenance items, and the establishment of a Bayesian Network model for vehicle quality evaluation. The vehicle design parameters, price and performance of functional systems are modeled as node variables in the Bayesian Network. Bayesian Network reasoning is then used to analyze the influence of these nodes on vehicle quality and their respective importance.

Findings – A case study using the maintenance records of 74 sport utility vehicle (SUV) models is presented to demonstrate the validity of the proposed framework. Our results reveal that factors such as vehicle size, chassis issues and engine displacement, can affect the chance of vehicle failures and accidents. The influence of factors such as price and performance of engine and chassis show explicit regional differences.

Originality/value – Previous research usually focuses on limited maintenance records from a single vehicle producer, while our proposed framework enables efficient and systematic processing of larger-scale maintenance records for vehicle quality analysis, which can support auto companies, consumers and regulators to make better decisions in purchase choice-making, vehicle design and market regulation.

Keywords Vehicle maintenance records, Text mining, Bayesian network, Vehicle quality evaluation

Paper type Research paper

Nomenclature

$\mathcal{N}$	Bayesian network	N	The number of total maintenance records for a vehicle model
$P(X_1, \ldots X_n)$	Joint probability distribution		
PPH	Number of quality problems per 100 vehicles	Q_1	A quality index that reflects the percentage of repair services in all services
R	Repair		
M	Maintenance		
A	Accident-related Repair	Q_2	A quality index that reflects the percentage of repair items in all services

The authors would like to acknowledge the financial support from the National Natural Science Foundation of China (52005328) and Shanghai Science and Technology Commission "Yangfan" Program (20YF1419300). The authors would also like to acknowledge the technical support from Mengyuan Shen and Bohan Feng.

International Journal of Quality & Reliability Management

0265-671X
DOI 10.1108/IJQRM-03-2023-0114

R_c	The number of the records that fall into the category of Repair	R_p	The number of repair items in N maintenance records
A_1	The number of the records that fall into the category of Accident-related repair per N maintenance records		

1. Introduction

Data is a key booster for the automotive industry in the era of Industry 4.0 (Wang *et al.*, 2016). Today massive amounts of human and machine-generated data have become accessible from which valuable insights can be extracted (Aslam *et al.*, 2021; Tao *et al.*, 2018). As an important component of vehicle's lifecycle, maintenance and repair is essential to keep vehicles in healthy working status in which a number of maintenance records can be generated (Ildarkhanov, 2023; Romanowski and Nagi, 2004; Wagner *et al.*, 2021). By mining these records, insights on vehicle quality, traffic accidents, failure patterns of automotive parts can be obtained. These insights can assist different stakeholders in the automotive industry (e.g. consumers, manufacturers, part and service suppliers and regulators) with purchase choice-making, improvement of vehicle design, quality control in manufacturing process, supply chain management, pricing of vehicle insurance, regulation of automotive market, etc (Lee *et al.*, 2008; Prashar, 2022; Viscusi and Evans, 2016). For example, Chang and Hwang (2013) collected and analyzed the maintenance records of Toyota service centers in Taiwan from 2004 to 2006. They found that brake jitter was the most common quality problem, and the main cause is the change of brake disc thickness due to wear and corrosion. They built a numerical model to simulate the change of disc thickness over time and predicted the frequency of jitter failures, which can help designers improve the design of brake disc. Hu *et al.* (2015) collected the maintenance data of FAW (First Automobile Works) -Volkswagen Sagitar from auto stores, and set up a statistical model based on Weibull distribution and maximum likelihood estimation to predict the demand of spare parts such as oil filter, which can be used to optimize the inventory management of spare parts.

However, past research mostly focus on analyzing the maintenance records of vehicles from a single producer (Pecht *et al.*, 1990), and few explore the value of massive maintenance records of vehicles from multiple producers. In fact, it is crucial for one automaker to know about the maintenance and repair status of the general market so that it can improve its products continuously and keep its market competitiveness. Traditionally, automakers mainly rely on customer surveys to understand the maintenance and repair of other companies' products, such as the Vehicle Dependability Study (VDS), a questionnaire-based survey developed by J.D. Power (Aguwa *et al.*, 2012). This survey examines the quality of automobiles with a 1.0–3.5 years ownership period. Other similar surveys include consumer reports (Aguwa *et al.*, 2012), Germany Technische Überwachung Verein (Jerz and Winterhalter, 2020) and China automobile product quality index report (Akdeniz Ar and Kara, 2014). Although it is useful to understand the vehicle quality by surveying consumers' experience and opinions, the subjectiveness of respondents and the limited sample size (e.g. J.D. Power only samples about 200 respondents for one vehicle model every year in the VDS of China (Lin *et al.*, 2018)) may undermine the representativity of the survey data to some extent.

With the rapid development of digital market regulations and Internet of things, today massive vehicle maintenance records after de-privacy processing can be directly collected from vehicle repair shops and dealerships (Zhang *et al.*, 2014). Compared to the maintenance data from a single producer or customer survey data, these maintenance records across multiple automakers are more objective with a much larger quantity (usually in millions per year per city). In addition, records collected from different regions can support the understanding of the spatial heterogeneity of vehicle quality, which is particularly helpful in demand forecasting and regionally differentiated marketing. For example, along with the fast urbanization, China's automotive market has shown strong regional characteristics (Bi *et al.*, 2021). The levels of economic development and traffic infrastructure construction, and the driving habits of consumers in a region can impact the vehicle maintenance status significantly (Guajardo *et al.*, 2016). These insights can be used to make more adaptable service strategies (e.g. determination of warranty duration, pricing of auto insurance). More importantly, these massive vehicle maintenance records are collected in real-time, which allows the vehicle quality patterns to be captured longitudinally (Spasic, 2022). These advantages will contribute to the development of data-driven design, manufacturing and service in the digital age. Considering large-scale maintenance records are mainly unstructured texts and have not been well utilized before, a systematic framework for processing and analyzing such data for vehicle quality analysis is needed.

In this paper, we propose an integrated data-driven framework that can efficiently process and analyze large-scale vehicle maintenance records across multiple automakers. Within this framework, we develop a comprehensive dictionary of maintenance items by extracting and classifying useful information from unstructured maintenance descriptions. This dictionary is then used in the classification of maintenance records, which provides the basis for vehicle quality analysis. We leverage Bayesian Network (BN) to model the relationship between vehicle attributes, patterns of maintenance data and quality of vehicles, and investigate what factors contribute to vehicle failures and accidents significantly. The proposed framework is validated in a case study by using the data from China's automotive market. Although this paper uses passenger vehicle data as a demonstration, the proposed framework can be extended to the data mining of other vehicles' maintenance records, such as electric vehicles or engineering vehicles.

The rest of the paper is organized as follows. Section 2 introduces the related work and technical background in vehicle quality evaluation and text mining. Section 3 explains the overall framework for mining massive vehicle maintenance records and the key techniques involved. Section 4 presents the background, datasets and data preprocessing of a case study. Section 5 provides the results of the case study including the establishment of the vehicle maintenance dictionary, classification of maintenance records, vehicle quality evaluation and diagnostic inference based on BN reasoning. Section 6 summarizes the results of the case study and their implications, and discusses the main contribution of this study with recommendations for future work.

2. Related work and technical background

In this section, the techniques of product quality evaluation, keywords extraction in text mining and BN are reviewed.

2.1 Methods for product quality evaluation

Evaluating and improving product quality is crucial for manufacturers to ensure the safety and satisfaction of consumers, design innovative products and keep market competitiveness (Spasic, 2022). Common methods to evaluate product quality include reliability testing and customer survey (Badri *et al.*, 1995; Zhang *et al.*, 2000). In a reliability testing, engineers test a product's performance in a simulated environment and then predict the future behavior throughout its life cycle (Elsayed, 2012). For automobiles, typical reliability testing includes road test, electronics test and crash test.

Since reliability testing is usually performed prior to product launch, it is limited in tracking the quality performance of the product in real use scenarios. In contrast, customer survey can get customers' feedback on the product quality after a period of usage. For example, J.D. power developed initial quality study (IQS) and VDS (automobile durability study) (Zheng *et al.*, 2019), which investigates the quality of vehicles with ownership of

2–6 months and 2.5–3.5 years, respectively. Based on these survey data, a popular metric, number of problems per 100 vehicles (PPH), is calculated to evaluate vehicle quality. It is defined as the average failures occurred in one hundred surveyed automobiles as shown in Equation (1),

$$PPH = \frac{P}{N} \times 100 \tag{1}$$

where P and N refer to the number of problems with vehicles and the number of vehicles participating in the survey, respectively, and lower PPH usually means better quality. J.D. Power provides a quality ranking of mainstream automobiles based on their PPH values every year. Consumers can leverage this ranking to make purchase decisions, and automakers can continuously improve their vehicle design by fixing the problems revealed in the survey. However, customer survey is subjective with limited sample size due to the high time and financial cost of conducting surveys. In this paper, we will explore how to get more objective insights from large-scale product maintenance records.

2.2 Keyword extraction methods

Keyword extraction is a key step in processing massive textual data such as vehicle maintenance records, and its performance directly affects the accuracy of classifying maintenance records and related vehicle quality analysis. Keyword extraction methods can be divided into supervised and unsupervised ones. In supervised methods, textual labels are needed to train a classifier (Rose *et al.*, 2010), and such labels are usually created manually with a high labor cost. Among unsupervised methods (Shah *et al.*, 2003), frequency-based methods and graph-based methods are widely used. We briefly review a representative method in each category.

2.2.1 TF-IDF (term frequency-inverse document frequency). TF-IDF is a traditional frequency-based keyword extraction method, which assumes that the importance of a word is proportional to the number of times it appears in a document (e.g. an article), but inversely proportional to the frequency of its occurrence in the whole corpus (i.e. a large-scale collection of well-sampled and processed texts, usually including millions or even billions of documents). Equations (2) to (4) present the basic idea of TF-IDF (Ramos, 2003),

$$TF_{w_i} = \frac{f_{w_i}}{F} \tag{2}$$

$$IDF_{w_i} = log\left(\frac{1}{DF_{w_i}}\right) = log\left(\frac{S}{S_{w_i}+1}\right) \tag{3}$$

$$TF - IDF_{w_i} = TF_{w_i} \times IDF_{w_i} \tag{4}$$

where f_{w_i} refers to the frequency of word w_i in a document, F represents the sum of the frequency of each word in the document. S is the number of all documents in the corpus. S_{w_i} refers to the number of documents containing w_i. The importance of w_i is then sorted by $TF{-}IDF_{w_i}$. However, the TF-IDF method does not consider the contextual and syntactic structure of words, and the importance of a word may not always match with its frequency (Hong and Zhen, 2012). Thus, TF-IDF is computationally efficient but sometimes performs unstably for certain datasets.

2.2.2 TextRank. TextRank is a popular graph-based keyword extraction method, in which an undirected text graph $G = (V, E, W)$ is constructed. Here $V = \{V_1, V_2, \ldots, V_n\}$ is a set of nodes representing n semantic units (i.e. words or phrases) in a text. $E = \{(V_i, V_j) | V_i, V_j \in V\}$ denotes the sets of edges representing the connectivity of the nodes (e.g. the syntactic relations between words) (Brin and Page, 1998). $W = \{w_{ij} | 1 \le i \le n, 1 \le j \le n\}$ is the set of edge weights, and w_{ij} is the weight of edge between node V_i and V_j. TextRank iteratively calculates the PR (Page Rank) value (i.e. weight) of each node base on PageRank algorithm (Kumbhar *et al.*, 2019), then the keywords can be obtained by sorting their PR values. Equation (5) shows this calculation (Rada Mihalcea, 1973):

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{5}$$

The sum on the right-hand side of Equation (5) represents the contribution of neighboring nodes to current node V_i, where $WS(V_i)$ is the PR value of node V_i, $In(V_i)$ refers to the set of nodes that point to V_i, and $Out(V_j)$ represents the number of nodes pointing from node V_j. d is a damping factor, which takes values in the range [0,1], indicating the probability of the current node jumping to any other node. w_{ij} indicates the similarity between nodes V_i and V_j. Considering its high compatibility of different textual formats, promising accuracy of keyword extraction and fast implementation (Firoozeh *et al.*, 2020), we choose TextRank as the main keyword extraction method in this research.

2.3 Bayesian network

BN is a graph to characterize the dependency relationships in complex systems (Langseth and Portinale, 2007). Let $\mathcal{N} = < G, P >$ be a BN, and G is an essentially directed acyclic graph (Li *et al.*, 2018). The values of the nodes in G are the random variables $X_1, X_2, \ldots, X_n$ and the directions of edges represent the causality or dependencies between nodes. P denotes the conditional probabilities of the nodes in G. In BN, if there is a link pointing from node X_i to node X_j, then X_i is the parent node of X_j, and X_j is the child node of X_i. The ancestors of a node include its parents and their ancestors. The root node has no ancestors. The descendants of a node include its children and their descendants. Each node is independent of its all nonparent nodes. For a node X_i, $P(X_i | \pi(X_i))$ represents the conditional probability of X_i at the occurrence of its parent $\pi(X_i)$. For a joint distribution of n random variables, the joint conditional probability of the network $P(X_1, \ldots X_n)$ is as follows (Li *et al.*, 2018):

$$P(X_1, \ldots X_n) = \prod_{i=1}^{n} P\left(X_i \middle| \pi(X_i)\right) \tag{6}$$

Equation (6) decomposes the BN into factorization forms with smaller local probability distributions, and each factorization form involves only one node and its parent. This is the basis for the subsequent calculation of the posterior probability of nodes. BN performs integrated reasoning by adjusting the conditional probability table of each node and synthesizing various prior knowledge for system analysis and evaluation.

BN has been used in previous studies on quality and reliability research. For example, Küçüker and Yet developed a BN model that can adjust and correct the product life estimates based on expert judgment and operational data in the reliability prediction of aircraft subsystems (Küçüker and Yet, 2022). The sum of squared error (SSE) between the expected predictive failure distribution and the actual time to failure is used as the performance metric. They found that BN can provide consistently accurate reliability predictions, and the performance of BN is better than data-driven parametric approach using maximum likelihood estimates (MLE) of the failure time modeled by the Weibull distribution or expert-based approach using mean time between failures (MTBF) values provided by the supplier. Ben Brahim *et al.* (2019) proposed to construct a BN based on an expert knowledge-based excitation method that integrates the occurring probability of the cause of failure and the detection information of potential failures before they occur from failure modes, effects and criticality analysis (FMECA). A case study using the thermoforming data of car floor covering was conducted to verify the applicability of the proposed method.

However, previous studies seldom combine the use of BN and the mining of massive maintenance records for vehicle quality analysis. In our research, we will diagnose and rank the key vehicle attributes that affect vehicle quality using BN based diagnostic inference. These results can support manufacturers to improve their design of products in a highly competitive market.

3. An integrated data-driven framework for vehicle quality analysis

3.1 Overview of the framework

Figure 1 shows the overall flow of the proposed data-driven framework for vehicle quality analysis based on massive vehicle maintenance records and BN. The rectangles represent data processing and modeling, and parallelograms represent the resulted data. The yellow color indicates the use of auxiliary data, the blue color indicates the use of vehicle maintenance records, and the pink color represents the obtained structured data. This framework consists of three major stages:

Stage 1: data collection. Vehicle maintenance records usually include the basic information of the vehicle (e.g. model, year, make and mileage) and descriptions on the mechanical inspection and maintenance service. However, these data will not be sufficient for comprehensive vehicle quality analysis. For example, detailed technical specifications of a vehicle model, such as its price, fuel consumption, size, weight, type of transmission and fuel, are usually not available in maintenance records but quite important when evaluating the influential factors of vehicle quality. We name these necessary data as auxiliary data. Other examples such as vehicle sensor data indicating the operating status of vehicles, can support the analysis of vehicle failures and accidents. Industrial or national standards on automobiles are useful to classify maintenance records in a

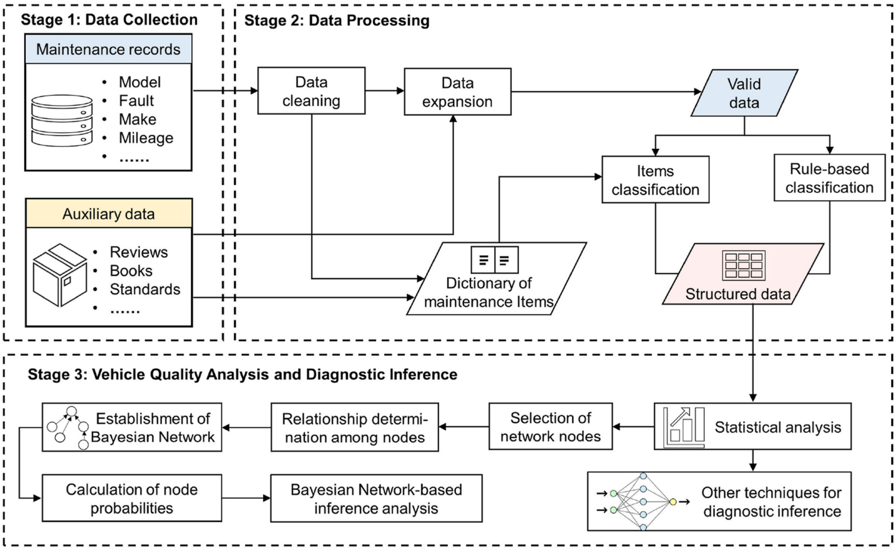


Figure 1.
The overall flow of the proposed framework

Source(s): Authors own work

systematic way. Thus, it is important to refer to specialized knowledge bases, websites and publications to collect necessary auxiliary data in Stage 1.

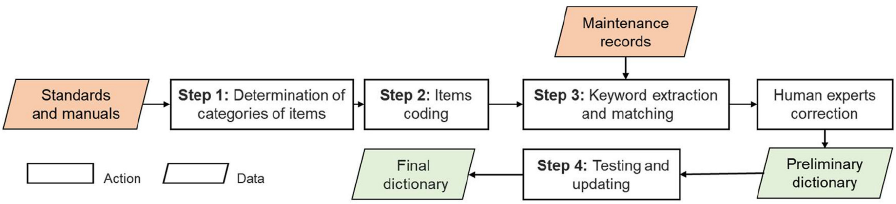
Stage 2: data processing. Data processing refers to converting raw unstructured data into clean, complete and structured data easily accessible by computers. This stage includes data cleaning, data expansion and classification of maintenance records. In data cleaning, the redundant, erroneous, missing and duplicate records from raw data are eliminated. Data expansion refers to combining necessary auxiliary data such as the tech specs of vehicles with the maintenance records to form a complete dataset. Since vehicle maintenance records are often unstructured texts, it is necessary to classify them into organized categories facilitating efficient analysis. In this process, a dictionary for maintenance items is needed, then the maintenance records can be classified by matching them with defined categories in the dictionary.

Stage 3: vehicle quality analysis and diagnostic inference. After obtaining the structured data from previous stage, statistical analysis on certain vehicle maintenance and repair items can be performed to get a general understanding about vehicle quality. Then diagnostic inference models can be built to analyze the relationships between vehicle attributes and vehicle quality. For a BN based model, the first step is to identify important vehicle attributes (e.g. vehicle design parameters, price) and the failure rates of major vehicle subsystems as nodes of the BN. Then the network structure of these nodes is determined by functional decomposition of vehicle system as well as regression analysis between vehicle attributes and vehicle quality performance obtained from the maintenance records. After that, a hierarchical BN can be formed, which can be used to infer the influence of these nodes on vehicle quality performance. Note that the obtained data can also be exported to other methods such as machine learning for diagnostic inference analysis.

To support the data processing and analysis in the proposed framework, we develop a comprehensive dictionary of vehicle maintenance items, propose a workflow for classifying maintenance records, design two vehicle quality indices and build a BN based model for vehicle quality analysis. The details of these techniques are presented in the following subsections.

3.2 The development of a comprehensive dictionary of vehicle maintenance items

A typical vehicle maintenance record often involves multiple maintenance items (e.g. mechanical inspection, change of faulty parts, change of oil and filter, etc.). The classification of maintenance items can reveal the associated technical problems with the serviced vehicle. To classify these items, a dictionary mapping them into corresponding categories is needed. Figure 2 illustrates the process of building the dictionary, in which the rectangles represent data processing and parallelograms represent the involved data. The orange color indicates the use of professional references and maintenance records, and the green color indicates the resulted dictionary. The details of each step are explained as follows.



Source(s): Authors own work

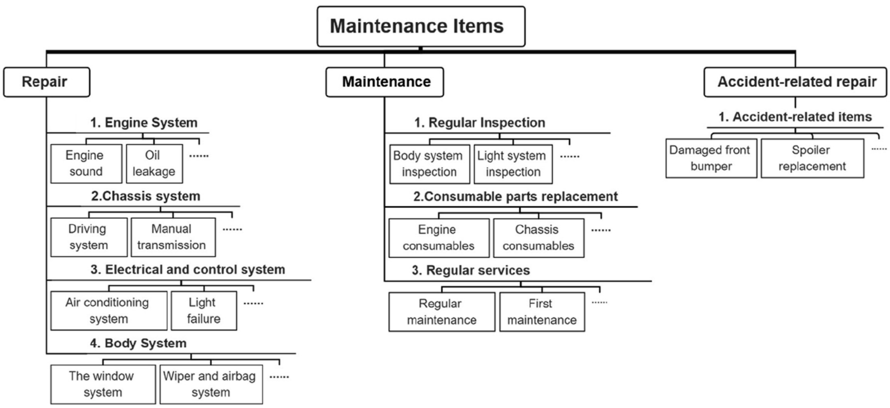
Figure 2. The process of building a comprehensive dictionary of vehicle maintenance items

Step 1: determination of the categories of maintenance items. In this research, we divide vehicle maintenance items into three basic categories: repair (items related to vehicle part faults, such as repairing the engine air intake system due to natural aging), maintenance (regular and necessary maintenance work, such as replacing oil filters) and accident-related repair (such as repairing the bumper damage caused by collisions). We then classify the maintenance items of each basic category into subcategories. For example, the repair items can be categorized into four main systems including engine, chassis, body, electrical and control. These systems can be further divided into subsystems according to functions, such as the subsystems of ignition, lubricating, air intake and fuel in the engine system. We collect these categories from vehicle manuals and professional references (Hirsch, 2014). All maintenance items are eventually organized in a tree structure as shown in Figure 3.

Step 2: item coding. After the categories of maintenance items are determined, a five-digit code is assigned to each category. For example, oxygen sensor failure is coded as 11,108. The first three digits 111 represent the repair (category), engine system (main system), engine sensor (subsystem), respectively. The last two digits 08 denote that this failure is the eighth item in the subsystem of engine sensor. These codes will smooth further data analysis.

Step 3: keyword extraction and matching. Each entry in the dictionary consists of a category (e.g. Repair) and mapped keywords (e.g. cylinder, pressure and relay) so that maintenance records and items can be classified. As mentioned in Section. 2.2, the TextRank method is used to extract keywords from the raw records, particularly those descriptions on vehicle faults and maintenance items. The Skip-Gram model (Liu *et al.*, 2015) is used to embed these keywords into low-dimensional word vectors. These keywords are then sorted by their occurring frequencies, and those high-frequency keywords are mapped with the entries in the dictionary (i.e. the identified categories) by calculating the similarity between word vectors (Jin *et al.*, 2018) as shown in Equation (7),

$$Sim_C(i,j) = \frac{\overrightarrow{A_i}.\overrightarrow{A_j}}{|\overrightarrow{A_i}| \times |\overrightarrow{A_j}|} \tag{7}$$



Source(s): Authors own work

Figure 3.
The tree structure of organized vehicle maintenance items

where $\overrightarrow{A_i} = \{A_1, A_2, A_3, \ldots, A_n\}$ is the word vector of word *i*, *n* is the vector size. A popular choice of *n* is 50 considering the number of the words in maintenance records is not too large (Jatnika *et al.*, 2019). $|\overrightarrow{A_i}|$ represents the length of $\overrightarrow{A_i}$. We keep those high-frequency keywords whose similarity with the keyword i in an entry of the dictionary exceeds a threshold value (usually 0.5). These kept keywords are then manually screened by human experts, and a preliminary dictionary of vehicle maintenance items is obtained.

Step 4: testing and updating. The validity and completeness of the dictionary is tested repeatedly using datasets randomly sampled from the raw maintenance records. Those missing items or items that do not match with the keywords in the maintenance records will be updated iteratively and the updated dictionary will be used in further analysis.

3.3 Classification of vehicle maintenance records and vehicle quality indices

The most important step in generating structured maintenance record data is to classify the maintenance records and items into defined categories. We propose a workflow for this classification based on dictionary matching (Cole *et al.*, 2004) and defined rules as shown in Figure 4. Here blue parallelograms represent the input data, yellow parallelogram indicates the obtained records with matching items, and the pink parallelogram represents the final obtained structured data.

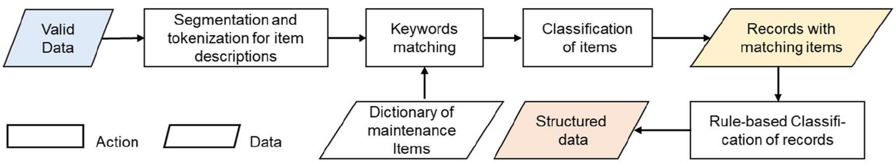
We first divide the descriptions on maintenance items into phrases or words using the segmentation and tokenization method from the "jieba" library of Python (Day and Lee, 2016). When a set of phrases or words from one maintenance item contains all the keywords of an entry in the established dictionary, this item is then matched with the corresponding category. Note that there could be multiple items within one maintenance record, thus the above process will be iterated through all items of all records. After finishing the classification of maintenance items, we further classify the maintenance records into three aggregated categories, namely repair (labeled as R), maintenance (labeled as M) and accident-related repair (labeled as A), which can explicitly indicate the major type of a vehicle maintenance service. If a maintenance record involves multiple items (e.g. a car receives regular maintenance after repair due to part faults), the following classification rules will be applied:

(1) If one or more accident-related repair items appear in the record, this record gets a label of A.

(2) If only maintenance items appear in the record, this record gets a label of M.

(3) If one or more repair items appear and no accident-related items exist in the record, this record gets a label of R.

The obtained data containing category codes for maintenance items and classification labels for maintenance records are merged as the final structured data, which can support the calculation of two indices to evaluate vehicle quality from a holistic view. The quality index Q_1 is calculated as shown in Equation (8),



Source(s): Authors own work

Figure 4. Workflow for the classification of maintenance records and items

$$Q_1 = R_c/N \tag{8}$$

where R_c is the number of the records that fall into the category of repair, and N is number of total maintenance records for a vehicle model during a period of time. This index reflects the percentage of repair services in all services, and a smaller value may imply less need of repair service and better quality of the vehicle model. Quality index Q_2 is calculated as shown in Equation (9),

$$Q_2 = R_p/N \tag{9}$$

where R_p is the number of repair items in N records. Since one maintenance record often includes multiple maintenance items, R_p is usually much larger than its corresponding R_c in a given dataset. This index is comparable to the index of *PPH* proposed by J.D. power, and a smaller Q_2 implies a lower number of failures and possibly a better vehicle quality.

3.4 A Bayesian Network based model for vehicle quality analysis

As a method for characterizing the dependency relationships in complex systems, BN can better describe the uncertainties involved in product or system quality diagnosis compared to other methods such as partial least squares (PLS) and principal component analysis (PCA) (Uusitalo, 2007). We propose a BN based model to describe the dependencies between vehicle quality and vehicle attributes, and probabilistic inferences can be then used to identify those most influential vehicle attributes. The establishment of the BN model includes the determination of nodes, network structure and the calculation of the probabilities with respect to each state of nodes. The detailed procedures are as follows.

3.4.1 Determination of nodes and network structure. According to previous research on vehicle design (O'Neill, 2009; Shaheen and Niemeier, 2001), vehicle design parameters, performance of vehicle functional systems, and vehicle quality metric are defined as the network nodes. Then the network structure can be determined by the dependencies among these nodes. Our preliminary study shows that vehicle design parameters such as height, width, engine displacement can impact the probability of vehicle accidents. In addition, the performance of main functional systems (i.e. engine, chassis, body, electronics and control) can influence the failure patterns of the vehicle. Based on these results, we divide the network nodes into three categories: root node (node without parent node), leaf node (node without child node), and descendant node (node with both parent and child nodes). Specifically, the root nodes include product design parameters, make origin and performance of vehicle subsystems. The performance of major functional systems, accident rate and failure rate as well as the price of the vehicle are defined as descendant nodes. The only leaf node is the vehicle quality metric. After that, the topology of the BN is established. The basic assumption is that the root nodes of the vehicle (e.g. design parameters, performance of subsystems, make origin) affects the performance of descendant nodes (e.g. performance of major functional systems), and the descendant nodes eventually affect the leaf node (i.e. vehicle quality).

3.4.2 Determination of node states. In order to get probabilistic inference from a BN, the node states need to be determined and discretized for computational convenience. For root and descendant nodes, their states are discretized into 0 and 1. For nodes such as performance of subsystems, design parameters and make origin, their states are determined by cutoff points. The cutoff point of the performance of subsystem nodes is the average number of repair items for a subsystem in certain amount of maintenance records. The cutoff points of design parameters are their mean values from the market data. Then the state of a node below the cutoff point is set to 0, otherwise 1. Apart from that, the state of make origin node is set to 1 if the vehicle owns a domestic brand, otherwise 0.

For the leaf node (vehicle quality), four states are assigned, which are determined by the combination of two states of two indicators (low and high of failure rate Q_1 and accident rate A_1). Based on the Q_1 and A_1 of one vehicle model, its vehicle quality node can be classified into four states (0, 1, 2, 3), and the classification criteria and results are shown in Table 1.

3.4.3 Calculation of the probability of node. The probabilities to calculate in a BN include prior, conditional and posterior probability. Prior probability refers to the predicted probability before the occurrence of an event. Conditional probability is the probability of one event occurring given that another event has occurred. Posterior probability is the probability that an event is caused by a certain factor given that this event has occurred. In our research, the prior and conditional probabilities are calculated from the statistical analysis of maintenance records. The prior probability of a root node in BN is obtained by calculating the ratio between the maintenance records of the vehicle in one state ($X_i \in \{0, 1\}$) and the total number N of maintenance records of the sample as shown in Equation (10):

$$P(X_i = k) = \frac{m}{N} \tag{10}$$

Here $k = 0$ or 1, representing the node state, and m is the number of maintenance records corresponding to the node state k. The conditional probabilities are determined by the conditional probability formula as shown in Equation (11):

$$P(A|B) = P(AB)/P(B) \tag{11}$$

Here A and B are two nodes with states. However, in some cases, the conditional probabilities may not be available. For example, suppose we want to calculate the conditional probability $P(C = 1|A = 1, B = 0)$, where A, B and C denote the body length, weight and price of the vehicle model, respectively (their state is 0 if the node variable is below the mean, otherwise 1). In fact, there may not be a vehicle model with $A = 1$ and $B = 0$, i.e. its body length is above the mean but the weight is below the mean. In this situation, we use the weighted average of the probabilities of other states that are similar to the state of interest as its interpolated conditional probability. Here the similarity is calculated based on the Euclidean distance between the state of interest and the other states with available conditional probabilities (Xiong and Yao, 2021). Figure 5 shows an example of this treatment. Based on similarity measurements, the most similar cases to $A = 1, B = 0$ are $A = 0, B = 0$ and $A = 1, B = 1$. Then, the mean of the probabilities of these two cases is taken as the missing probability.

The posterior probability derived from BN inference is based on variable elimination method (Chavira and Darwiche, 2007) as shown in Algorithm 1. This method converts the global probability inference calculation into the factor product and summation operation of local node variables through the existing joint probability distribution $P(X_1, \ldots X_n)$. It gradually eliminates the other node variables unrelated to the evidence variables (i.e. node

Table 1. Description of four vehicle quality states

Vehicle quality state	Failure rate	Accident rate
0	Low ($Q_1 < \overline{Q_1}$)	Low ($A_1 < \overline{A_1}$)
1	Low ($Q_1 < \overline{Q_1}$)	High ($A_1 \geq \overline{A_1}$)
2	High ($Q_1 \geq \overline{Q_1}$)	Low ($A_1 < \overline{A_1}$)
3	High ($Q_1 \geq \overline{Q_1}$)	High ($A_1 \geq \overline{A_1}$)

Note(s): Q_1 and A_1 are the number of the records that fall into the category of Repair or Accident-related repair per N maintenance records, respectively. $\overline{Q_1}$ and $\overline{A_1}$ are the average of failure rate Q_1 or average of accident rate A_1 with multiple sampling from the raw dataset, respectively
Source(s): Authors own work

variables with known probabilities) and query variables (i.e. node variables that need to calculate the posterior probabilities) in a specific order (Chavira and Darwiche, 2007). Then, the posterior probability distribution of the query variable can be obtained.

Algorithm 1. Variable elimination: $VE(\mathcal{N}, \boldsymbol{E}, \boldsymbol{e}, \boldsymbol{Q}, \rho)$

Input: The Bayesian Network $\mathcal{N}$. Evidence variable $\boldsymbol{E}$. The value of observed variable e. $\boldsymbol{Q}$ is the query variable. ρ is the order of elimination.

Output: Posterior probability distribution $P(\boldsymbol{Q} \mid \boldsymbol{E} = \boldsymbol{e})$.

$F \leftarrow$ The set of all probability distributions in $\mathcal{N}$.
$E \leftarrow e$
while $\rho \neq \emptyset$ *do*
  Z be the top variable in ρ
  Z is removed from ρ
  $F \leftarrow \mathrm{Elim}(F, Z)$
end while
$h(\boldsymbol{Q}) \leftarrow$ Multiply all factors in F.
return $P(\boldsymbol{Q}|\boldsymbol{E} = \boldsymbol{e}) = h(\boldsymbol{Q}) / \sum_Q h(\boldsymbol{Q})$

Elim (F, Z)

Input: F is the set of all probability distributions in the network. Z is the variable to be eliminated.

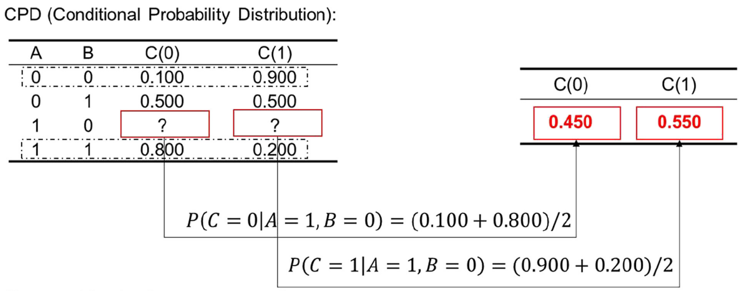
Output: Another set of functions.

Remove $\{f1, f2, \dots, fk\}$ involving the Z from F
$g \leftarrow \prod_{i=1}^{k} fi$
$h \leftarrow \sum_Z g$
Put h back into the F
return F

4. A case study of passenger vehicles

To demonstrate the validity of the proposed framework, we present a case study employing the vehicle maintenance records collected by a data service company. The dataset contains around 20 million vehicle maintenance records from auto dealerships in each of two provincial administrative regions of China (Shanghai and Shandong) between 2015 and 2020. These records include vehicle-related attributes (e.g. mileage, model) and maintenance information (e.g. record ID (identifier), fault description, repair date, item description, repair parts). The privacy-related information had been removed when we received the dataset. Table 2 shows a sample of the raw dataset after data cleaning.

In this case study, we apply the proposed framework to investigate what factors are mostly related to vehicle quality, how these factors influence quality and examine whether



Source(s): Authors own work

Figure 5.
An example of missing probability interpolation

such influences will vary between regions. Particularly, we are interested in Sport Utility Vehicles (SUVs), which is a vigorous segment in China' automobile market (Jiang *et al.*, 2023; Wang *et al.*, 2015). We randomly sampled 5,000 maintenance records for each of the 74 best-selling SUV models from both Shandong and Shanghai (i.e. 370 thousand records for each region), whose sales account for more than 75% of the whole SUV market.

The descriptive statistics of these SUVs from the sampled datasets are summarized in Table 3. The continuous variables include *Vehicle Price* (RMB, the currency unit of China), *Engine Displacement* (milliliter), and *Repair Mileage* (kilometer). Categorical variable includes *Make Origin* showing the origin of the car brand, which is an intrinsic attribute regardless of the region. It can be seen that the average *Price*, *Displacement*, *Mileage* of SUVs in Shanghai are slightly higher than those in Shandong. European and American SUVs hold the largest market share, followed by Chinese and Japanese SUVs.

5. Results

In this section, we first present the established dictionary and the classification results of maintenance records and items. Then we show the built BN based model for vehicle quality analysis. Finally, we report the relationship between vehicle attributes and vehicle quality based on BN inference, and discuss the regional differences.

5.1 The established dictionary and classification of maintenance records

Table 4 shows sample entries of the established dictionary of maintenance items. A typical entry in the dictionary includes item code, item name, associated keywords, number of keywords, and related vehicle subsystem, main system and main category. For example, "Fuel Pressure Regulating Valve Failure", coded as 11508, is a repair item under the fuel supply subsystem and the engine system. Our built dictionary includes 451 maintenance

Record ID	Model	Mileage	Fault description	Repair date	Item description	Repair parts
65752012	Buick Envision	32,321 km	"Add fuel additive"	2016/4/18	"Add fuel additives to new cars"	"Gasoline detergent 150 ML"

Source(s): Authors own work

Table 2. A sample of the raw dataset

Continuous variables	Shandong Mean	Shandong Min	Shandong Max	Shanghai Mean	Shanghai Min	Shanghai Max
Vehicle price (RMB)	254,340	70,900	639,900	268,230	70,900	653,000
Engine displacement (milliliter)	1,925	999	3,778	1,952	1,299	3,778
Repair mileage (Kilometer)	34,141	101	588,888	36,614	101	594,900

Categorical variable Make origin	Chinese (%)	America (%)	Japan (%)	Korea (%)	Europe (%)
	75,000	80,000	75,000	20,000	120,000
	20.27%	21.62%	20.27%	5.41%	32.43%

Source(s): Authors own work

Table 3. Descriptive statistics of SUVs from the sampled datasets

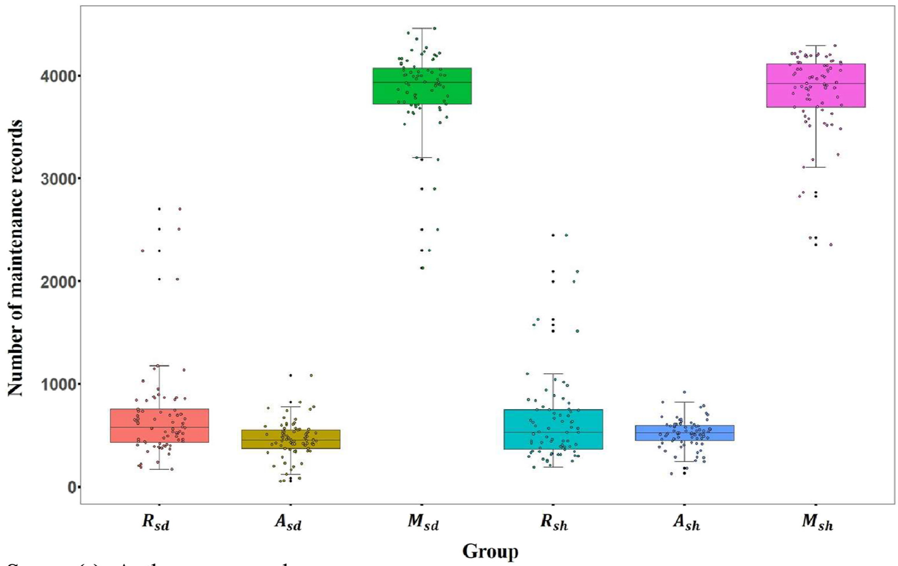
items, which belong to 29 subsystems, 7 main systems and 3 main categories. The average number of associated keywords for each maintenance item is 1.64.

Figure 6 presents the classification of maintenance records for SUVs in Shandong and Shanghai using the dataset introduced in Section 4. A dot in the group R_{sd} represents the number of maintenance records of one SUV model classified as Repair (R) in Shandong, and the corresponding box plot shows the four quantiles of the distribution. The interpretations of the other symbols are alike. We can observe that the records falling into the category of *Maintenance* account for about 70–80% of total records in each region, which indicates that the services in dealerships are dominated by routine maintenance. When comparing the differences between two regions, we find that the mean number of *Repair* records in Shanghai ($\mu = 646, std = 438$) is lower than that in Shandong ($\mu = 690, std = 468$), though this difference is not significant in a two-sample t-test ($t = 0.602, p = 0.549$). In addition, there are some scattered points that deviate far from the mean of boxplots, which implies that these SUV models may have more quality problems compared to average.

Table 4.
Sample entries from the dictionary of maintenance items. R, M and a indicate the items falling into the category of repair (R), maintenance (M) or accident-related repair (A)

Code	Item	Keywords	Number of keywords	Subsystem	Main system	Category
11508	Fuel Pressure Regulating Valve Failure	Fuel, pressure, Regulating	3	Fuel supply	Engine	R
22112	Engine cover inspection lubrication	Engine, inspection, lubrication	3	Engine inspection	Inspection	M
30027	Water tank frame shaping	Water tank, frame	2	–	–	A

Source(s): Authors own work



Source(s): Authors own work

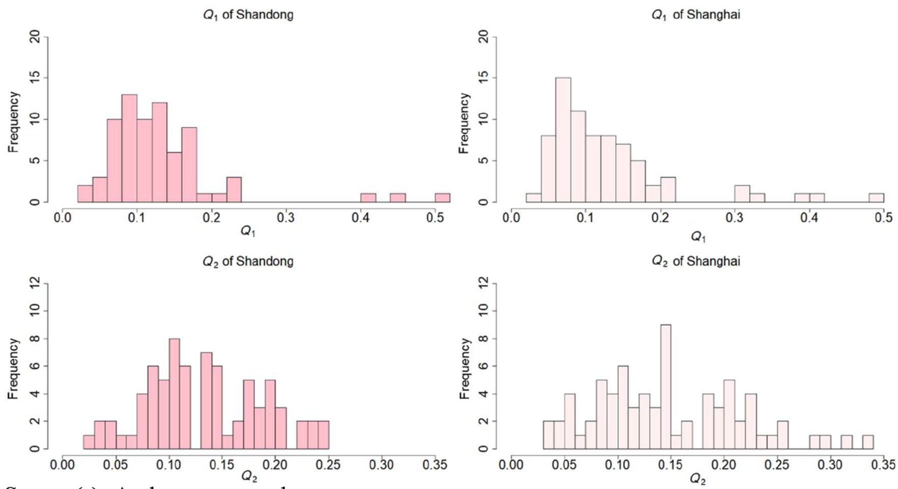
Figure 6.
Classification results of maintenance records for SUVs in Shandong (with the subscript of sd) and Shanghai (with the subscript of sh)

After getting the classification of maintenance records, we calculated two quality indices of 74 SUV models using Equations (8)-(9). Their distributions are shown in Figure 7 smaller Q_1 or Q_2 indicates better performance in the quality evaluation. Our results show that Q_1 and Q_2 of most SUV models concentrate around 0.100 to 0.150. Additionally, SUVs in Shanghai may have better performance than SUVs in Shandong as the former have a lower mean of Q_1 ($\mu = 0.129, std = 0.088$) than the latter ($\mu = 0.138, std = 0.094$). Nevertheless, the Q_2 of some SUV models in Shanghai exceeds 0.25 or even 0.3 and the average of Q_2 in Shanghai ($\mu = 0.148, std = 0.069$) is larger than that in Shandong ($\mu = 0.133, std = 0.054$), which could be resulted from that a maintenance record may averagely include more maintenance items for some SUV models in Shanghai than that in Shandong. This result reveals that the quality index Q_1 can reflect the frequency of vehicle failures, which is more suitable for measuring the overall quality of the vehicle, while Q_2 is better to measure the cost required for fixing these failures.

5.2 The Bayesian Network model for vehicle quality inference

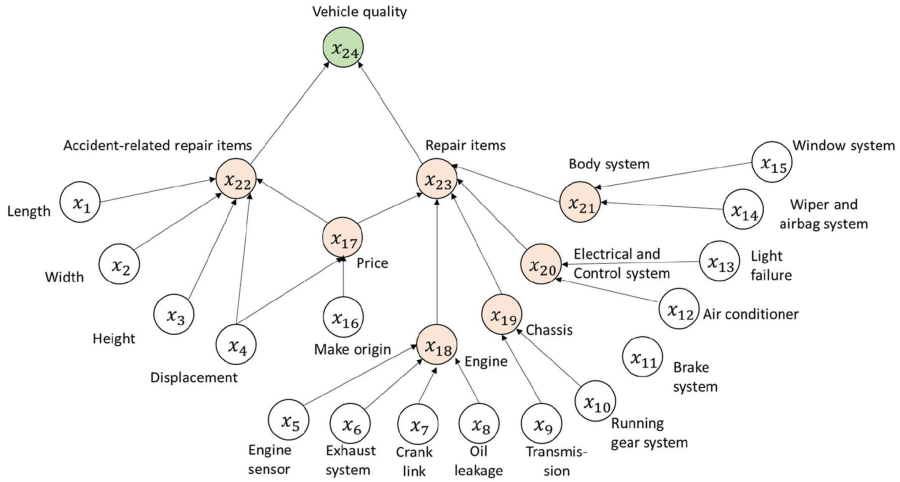
To investigate what factors are mostly related to vehicle quality and how these factors influence quality, we built a BN model as shown in Figure 8. This network includes 16 root nodes (x_1 to x_{16}), 7 descendant nodes (x_{17} to x_{23}) and 1 leaf node (x_{24}). The root nodes include vehicle design parameters (e.g. length, width, height), and the performance of basic functional subsystems of vehicles. The descendant nodes reflect the performance of higher-level functional systems and price. The leaf node indicates the overall quality evaluation of vehicle. Such network allows us to explore how vehicle design parameters and the performance of functional systems influence the overall quality of vehicles.

As mentioned in Section. 3.4, BN based diagnostic inference needs the calculation of prior and conditional probabilities of nodes, and a reasonable setting of the node states and cutoff value of each state is the first step. Table 5 presents the nodes and their states with cutoff values in the built BN. For example, if the body length of a SUV model is less than 4,600 mm, then the state of x_1 (Length) is set to 0, otherwise 1. The state of x_{16} (Make Origin) is set to 1 for vehicles with domestic brands and 0 for others. The leaf node x_{24} (Vehicle quality) has four states, determined by the combination of two states of two indicators (low and high of failure rate Q_1 and accident rate A_1).



Source(s): Authors own work

Figure 7. Histograms of quality indices Q_1 and Q_2 for SUVs in Shandong and Shanghai



Source(s): Authors own work

Figure 8.
The structure of the Bayesian network model

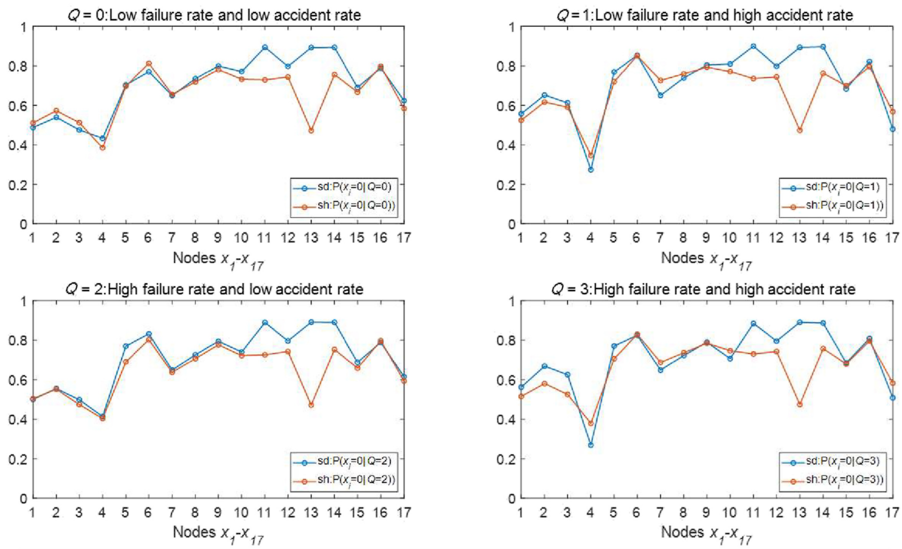
Category of nodes	Node	Definition	States of nodes with cutoff values	
Design parameters (mm)	x_1	Length	0 (<4,600)	1 (≥4,600)
	x_2	Width	0 (<1,850)	1 (≥1,850)
	x_3	Height	0 (<1,680)	1 (≥1,680)
Design parameters (milliliter)	x_4	Displacement	0 (<1,800)	1 (≥1,800)
Subsystem (items/1,000 records)	x_5	Engine sensor	0 (<8)	1 (≥8)
	x_6	Exhaust system	0 (<5)	1 (≥5)
	x_7	Crank link	0 (<24)	1 (≥24)
	x_8	Oil leakage	0 (<9)	1 (≥9)
	x_9	Transmission	0 (<8)	1 (≥8)
	x_{10}	Running gear system	0 (<39)	1 (≥39)
	x_{11}	Brake system	0 (<4)	1 (≥4)
	x_{12}	Air conditioner	0 (<11)	1 (≥11)
	x_{13}	Light failure	0 (<6)	1 (≥6)
	x_{14}	Wiper and airbag system	0 (<5)	1 (≥5)
	x_{15}	Window system	0 (<16)	1 (≥16)
Make Origin	x_{16}	Make Origin	0 (Overseas)	1 (Domestic)
Price (RMB)	x_{17}	Price	0 (<250,000)	1 (≥250,000)
Main system (related records/ 1,000records)	x_{18}	Engine	0 (<210)	1 (≥210)
	x_{19}	Chassis	0 (<47)	1 (≥47)
	x_{20}	Electrical and control system	0 (<18)	1 (≥18)
	x_{21}	Body system	0 (<17)	1 (≥17)
Number of items (items/1,000records)	x_{22}	Accident-related repair items	0 (<168)	1 (≥168)
	x_{23}	Repair items	0 (<513)	1 (≥513)
Vehicle quality	x_{24}	Vehicle quality	0 ($Q_1 < 0.109$; $A_1 < 0.104$) 1 ($Q_1 < 0.109$; $A_1 \geq 0.104$) 2 ($Q_1 \geq 0.109$; $A_1 < 0.104$) 3 ($Q_1 \geq 0.109$; $A_1 \geq 0.104$)	

Source(s): Authors own work

Table 5.
Nodes with their states in the Bayesian network

After determining the states of all nodes, the prior probabilities of each root node and the conditional probabilities of other nodes can be calculated as explained in Section 3.4.3. Then the posterior probabilities of root nodes can be obtained. Figure 9 shows the posterior probabilities of all root nodes x_1 to x_{16} and node x_{17} (Price) for a given vehicle quality state ($Q = 0, 1, 2, 3$) in Shandong and Shanghai. These nodes are mostly basic attributes of vehicles. Take the x_{17} (Price) in Shandong (see the orange dots) as an example, if a SUV model's vehicle quality falls into the category of low failure rate and low accident rate (i.e. $Q = 0$), then the probability that its price is less than market average (i.e. $P(x_{17} = 0|Q = 0)$) is 0.624. Similarly, if the vehicle quality falls into the category of low failure rate and high accident rate (i.e. $Q = 1$), then such probability (i.e. $P(x_{17} = 0|Q = 1)$) becomes 0.480. This result implies that the increase of vehicle accident rate may relate to the increase of vehicle price. A possible explanation is that pricier SUVs are usually larger in size and vulnerable to scuffing and scraping (Takubo and Mizuno, 2000). In addition, the owners of these vehicles are more inclined to go to brand dealerships for maintenance rather than small repair shops with less data reported.

As can be seen from Figure 9, the posterior probabilities of most nodes do not change equally for the four states of vehicle quality, such as design parameter x_4 (displacement) and functional system of chassis x_{10} (Running gear system). In contrast, the posterior probability of x_{17} (Price) seems to equally change across different states of vehicle quality, which indicates that the quality of vehicles in Shandong may have potential relationship with price. However, this observation does not apply to Shanghai, which may be related to the different levels of local economic development, since Shanghai's per capita GDP (Gross Domestic



1: Length	4: Displacement	7: Crank link	10: Running gear system	13: Light failure	16: Make Origin
2: Width	5: Engine sensor	8: Oil leakages	11: Brake system	14: Wiper and airbag system	17: Price
3: Height	6: Exhaust system	9: Transmission	12: Air conditioner	15: Window system	

Source(s): Authors own work

Figure 9. Posterior Probabilities of 16 root nodes and price node for four states of the leaf node (vehicle quality) in Shandong (sd) and Shanghai (sh). The numbers 1,2,3 . . . 17 on the horizontal axis correspond to the nodes $x_1, x_2, x_3 \dots x_{17}$ respectively

Product) is 173,800 RMB and Shandong's is 81,800 RMB in 2021 (Pan *et al.*, 2022). For those attributes related to electrical and control system and body system, such as x_{13} (Light failure) and x_{15} (Window system), their posterior probabilities do not differ significantly for different states of vehicle quality, and this finding applies in both Shandong and Shanghai. This result implies that the performance of electrical and control and body system of these SUV models are relatively stable in the two regions.

5.3 Assessment of the influence and importance of different vehicle attributes on vehicle quality

According to Table 5, vehicle quality (x_{24}) includes four states ($Q \in \{0, 1, 2, 3\}$), which corresponds to the combination of low and high of failure rate and accident rate. Table 6 shows the nodes with the largest change of posterior probability during the evolution of vehicle quality states in Shandong and Shanghai. A large change may imply strong influence of the node on vehicle quality. For example, in Shandong, the posterior probability of node x_4 (Displacement) is smaller than market average given that the vehicle has both low failure rate and accident rate ($P(x_4 = 0|Q = 0)$) is 0.434, while the same probability under that a vehicle has low failure rate but high accident rate ($P(x_4 = 0|Q = 1)$) is 0.274. Here the difference implies that the engine displacement may have some influence on vehicle accident rate. This influence can also be captured by calculating the change of such probabilities during the evolution from $(Q = 2)$ to $(Q = 3)$ where the failure rates are both high but the accident rate increases. Similarly, the change of posterior probabilities during the evolution from $(Q = 0)$ to $(Q = 1)$ and from $(Q = 1)$ to $(Q = 3)$ can reflect the influence of vehicle attributes on vehicle failure rate.

When examining the evolution processes of $(Q = 0) \rightarrow (Q = 1)$ and $(Q = 2) \rightarrow (Q = 3)$ in Table 6, we find x_3 (Height) and x_2 (Width) may have potential influence on vehicle accident rate. SUV models with smaller vehicle height or weight can be more sensitive to complex driving conditions such as bumpy roads and lower stability and passability may lead to a higher probability of vehicle accidents. In addition to these design parameters, the accident rate of the SUV models in Shandong is greatly affected by x_{17} (Price), but this observation does not apply for Shanghai.

When examining the evolution processes of $(Q = 0) \rightarrow (Q = 2)$ and $(Q = 1) \rightarrow (Q = 3)$ in Table 6, we find the posterior probabilities of x_5 (Engine sensor), x_6 (Exhaust system), x_7 (Crank Link) under the engine system and x_{10} (Running gear system) change the most and could influence the vehicle failure rate greatly. These four systems are critical to the normal operation of vehicles and failed vehicles usually have large number of repair items related to them. In addition, we find the failure rate of SUV models in Shanghai seems to be more related with the vehicle design parameters, while the performance of engine system has a greater impact on vehicle failure rate in Shandong. One possible explanation is that vehicle owners in Shanghai are more willing to conduct routine maintenances, which can reduce the probability of engine failures (Ouyang *et al.*, 2019). The vehicle parts in Shandong are aging and wearing more seriously due to lower maintenance frequency.

In order to assess the importance of these vehicle attributes' influence on vehicle quality, we suggest two indices, Imp_{Ai} and Imp_{Ri}, to denote the importance of the influence of node x_i on the vehicle accident rate and failure rate, respectively. Their calculations are shown in Equations (12) and (13).

$$\begin{aligned} Imp_{Ai} = {} & |P(x_i = 0|Q = 0) - P(x_i = 0|Q = 1)| + \\ & |P(x_i = 0|Q = 2) - P(x_i = 0|Q = 3)| \end{aligned} \tag{12}$$

$$\begin{aligned} Imp_{Ri} = {} & |P(x_i = 0|Q = 0) - P(x_i = 0|Q = 2)| + \\ & |P(x_i = 0|Q = 1) - P(x_i = 0|Q = 3)| \end{aligned} \tag{13}$$

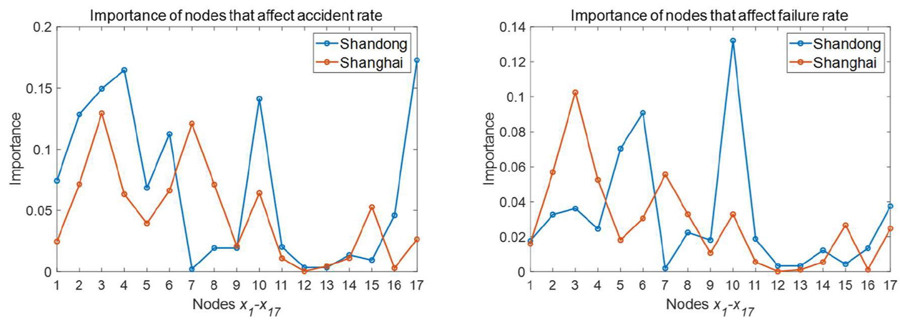
Table 6. Nodes with largest change of posterior probabilities during the evolution of vehicle quality states in Shandong and Shanghai

Evolutionary process	Shandong Node	Shandong Definition	Shandong ΔP	Shanghai Node	Shanghai Definition	Shanghai ΔP
$(\boldsymbol{Q}=\mathbf{0}) \rightarrow (\boldsymbol{Q}=\mathbf{1})$	x_4	Displacement	0.434 → 0.274 (−0.160)	x_3	Height	0.512 → 0.591 (0.079)
Same low failure rate	x_{17}	Price	0.624 → 0.480 (−0.144)	x_7	Crank link	0.655 → 0.725 (0.070)
Increased Accident rate	x_3	Height	0.476 → 0.612 (0.136)	x_2	Width	0.573 → 0.616 (0.043)
$(\boldsymbol{Q}=\mathbf{2}) \rightarrow (\boldsymbol{Q}=\mathbf{3})$	x_4	Displacement	0.414 → 0.269 (−0.145)	x_3	Height	0.475 → 0.526 (0.051)
Same high failure rate	x_3	Height	0.499 → 0.625 (0.126)	x_7	Crank link	0.637 → 0.688 (0.051)
Increased Accident rate	x_2	Width	0.556 → 0.668 (0.112)	x_8	Oil leakage	0.707 → 0.737 (0.030)
$(\boldsymbol{Q}=\mathbf{0}) \rightarrow (\boldsymbol{Q}=\mathbf{2})$	x_5	Engine sensor	0.703 → 0.771 (0.068)	x_3	Height	0.512 → 0.475 (−0.037)
Same low accident rate	x_6	Exhaust system	0.770 → 0.832 (0.062)	x_2	Width	0.573 → 0.552 (−0.021)
Increased Failure rate						
$(\boldsymbol{Q}=\mathbf{1}) \rightarrow (\boldsymbol{Q}=\mathbf{3})$	x_{10}	Running gear system	0.810 → 0.707 (−0.103)	x_3	Height	0.591 → 0.526 (−0.065)
Same high accident rate	x_{17}	Price	0.480 → 0.509 (0.029)	x_7	Crank link	0.725 → 0.688 (−0.037)
increased failure rate						

Note(s): ΔP is the change of the probabilities from one state to another and the change value is given in the parentheses
Source(s): Authors own work

Figure 10 presents the importance of the 17 nodes for vehicle accident rate and failure rate in Shandong and Shanghai. It can be seen that the nodes of the design parameters significantly affect the accident rate, and the most important nodes are x_3 (Height), x_4 (Displacement) and x_{17} (Price). When comparing the two regions, the importance of most nodes for vehicle accident rate are close to each other except x_7 (Crank link), x_{10} (Running gear system) and x_{17} (Price). In fact, the number of crank link related maintenance items of some SUV models in Shanghai are much higher than those in Shandong, and some of these SUV models have been recalled. In addition, the average price of SUV models in Shanghai ($\mu = 268,230$) is higher than that in Shandong ($\mu = 254,340$), which implies that customers in Shanghai may prefer higher-end versions of vehicles, while customers in Shandong may prefer normal ones. The differences of these models in performance and durability can also lead to the difference of price impact between the two regions.

When examining the importance of nodes on vehicle failure rate, we find the mechanical system nodes such as x_6 (Exhaust System), x_7 (Crank link) and x_{10} (Running gear system) are the most important ones. These nodes mainly belong to the engine system and chassis system with the highest number of repair items in our previous analysis. When comparing the regional differences, the importance of these nodes for vehicle failure rate varies greatly. One possible reason is that vehicle chassis systems are more prone to fail on lower-quality roads as they need to withstand the direct impact from ground. In addition to the level of economic development, natural factors such as temperature, can also affect the performance of engine intake and transmission systems (Culley *et al.*, 2009). The average annual temperature is 17.8 ℃ in Shanghai (Gu *et al.*, 2020) and 15.0 in Shandong (Zhang *et al.*, 2020) from 2014 to 2021, which may contribute to the difference of the importance of x_6 (exhaust system) and x_7 (crank link) between the two regions. According to these results, auto makers may consider allocating more resources to improve the design and manufacturing of running gear system and crank link in their next-generation SUVs.



1: Length	4: Displacement	7: Crank link	10: Running gear system	13: Light failure	16: Make Origin
2: Width	5: Engine sensor	8: Oil leakages	11: Brake system	14: Wiper and airbag system	17: Price
3: Height	6: Exhaust system	9: Transmission	12: Air conditioner	15: Window system	

Source(s): Authors own work

Figure 10.
Importance of 17 nodes on vehicle quality (accident rate and failure rate) in Shandong and Shanghai

6. Conclusion

In this paper, we propose an integrated data-driven framework for vehicle quality analysis from massive vehicle maintenance records. To support the data processing and analysis within the framework, a comprehensive dictionary on vehicle maintenance items, a classification workflow for maintenance records, and a BN model for diagnostic inference are developed. Our framework enables automatic processing of larger-scale vehicle maintenance records for vehicle quality analysis, and can support auto companies, consumers and regulators to make better decisions.

We demonstrate the validity of the proposed framework through a case study using data from two regions in China's automotive market. Our results show that the quality index that involves the counting of maintenance items instead of records can provide more detailed pictures on the quality of vehicles. The results from BN based diagnostic inference reveal that the quality problems of SUVs concentrate in engine and chassis systems, such as crank link and running gear system. Automobile manufacturers can increase their competitiveness by continuously fixing essential quality problems and improving their vehicle design in a timely manner from these insights. Auto suppliers can store parts in advance by analyzing the trend of vehicle quality problems. We also find that some vehicle design parameters (e.g. height) and performance of functional systems (e.g. crank link) significantly affect the accident rate as well as the failure rate of vehicles, and the influence and importance of these parameters vary between different regions. Automobile manufacturers and sellers can utilize these results to differentiate their design and marketing strategies correspondingly.

One limitation of this study is the incompleteness of the vehicle maintenance records. The collected records may not fully reflect the whole auto service market since only two regions' data are reported. We hope to conduct time-series analysis of more vehicle maintenance records in the future, which can reveal the spatial and temporal dynamics of vehicle quality, and support making more reasonable decisions in warranty strategy adjustment and vehicle design. Another limitation is that only linear relationships between vehicle quality and vehicle attributes are considered in the proposed framework. We expect to leverage nonlinear modeling techniques such as neural networks beyond BN for diagnostic inference analysis from these records.

References

Aguwa, C.C., Monplaisir, L. and Turgut, O. (2012), "Voice of the customer: customer satisfaction ratio based analysis", *Expert Systems with Applications*, Vol. 39 No. 11, pp. 10112-10119, doi: 10.1016/j.eswa.2012.02.071.

Akdeniz Ar, A. and Kara, A. (2014), "Emerging market consumers' country of production image, trust and quality perceptions of global brands made-in China", *Journal of Product and Brand Management*, Vol. 23 No. 7, pp. 491-503, doi: 10.1108/JPBM-12-2013-0472.

Aslam, B., Hu, J., Shahab, S., Ahmad, A., Saleem, M., Shah, S.S.A., Javed, M.S., Aslam, M.K., Hussain, S. and Hassan, M. (2021), "The nexus of industrialization, GDP per capita and CO2 emission in China", *Environmental Technology and Innovation*, Vol. 23, 101674, doi: 10.1016/j.eti.2021.101674.

Badri, M.A., Davis, D. and Davis, D. (1995), "A study of measuring the critical factors of quality management", *International Journal of Quality and Reliability Management*, Vol. 12 No. 2, pp. 36-53, doi: 10.1108/02656719510080604.

Ben Brahim, I., Addouche, S.A., El Mhamedi, A. and Boujelbene, Y. (2019), "Build a Bayesian network from FMECA in the production of automotive parts: diagnosis and prediction", *IFAC-PapersOnLine*, Vol. 52 No. 13, pp. 2572-2577, doi: 10.1016/j.ifacol.2019.11.594.

Bi, Y., Qiu, Y., Sha, Z., Wang, M., Fu, Y., Contractor, N. and Chen, W. (2021), "Modeling multi-year customers' considerations and choices in China's auto market using two-stage bipartite network analysis", *Networks and Spatial Economics, Networks and Spatial Economics*, Vol. 21 No. 2, pp. 365-385, doi: 10.1007/s11067-021-09526-9.

Brin, S. and Page, L. (1998), "The anatomy of a large-scale hypertextual Web search engine BT - computer Networks and ISDN Systems", *Computer Networks and ISDN Systems*, Vol. 30 Nos 1-7, pp. 107-117, doi: 10.1016/s0169-7552(98)00110-x.

Chang, Y.K. and Hwang, J.R. (2013), "Numerical model for prediction of brake judder due to wear and rust", *International Journal of Automotive Technology*, Vol. 14 No. 3, pp. 375-384, doi: 10.1007/s12239−013−0041−z.

Chavira, M. and Darwiche, A. (2007), "Compiling Bayesian networks using variable elimination", *IJCAI International Joint Conference on Artificial Intelligence*, pp. 2443-2449.

Cole, R., Gottlieb, L.A. and Lewenstein, M. (2004), "Dictionary matching and indexing with errors and don't cares", *Conference Proceedings of the Annual ACM Symposium on Theory of Computing*, pp. 91-100, doi: 10.1145/1007352.1007374.

Culley, D.E., Paluszewski, P.J., Smith, B.J. and Storey, W. (2009), "The case for distributed engine control in turbo-shaft engine systems", *Annual Forum Proceedings - AHS International*, Vol. 2 September, pp. 1732-1745.

Day, M.Y. and Lee, C.C. (2016), "Deep learning for financial sentiment analysis on finance news providers", *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, IEEE, Vol. 1, pp. 1127-1134, doi: 10.1109/ASONAM.2016.7752381.

Elsayed, E.A. (2012), "Overview of reliability testing", *IEEE Transactions on Reliability*, Vol. 61 No. 2, pp. 282-291, doi: 10.1109/TR.2012.2194190.

Firoozeh, N., Nazarenko, A., Alizon, F. and Daille, B. (2020), "Keyword extraction: issues and methods", *Natural Language Engineering*, Vol. 26 No. 3, pp. 259-291, doi: 10.1017/S1351324919000457.

Gu, Y., Li, K., Xu, J., Liao, H. and Zhou, G. (2020), "Observed dependence of surface ozone on increasing temperature in Shanghai, China", *Atmospheric Environment*, Vol. 221 November 2019, 117108, doi: 10.1016/j.atmosenv.2019.117108.

Guajardo, J.A., Cohen, M.A. and Netessine, S. (2016), "Service competition and product quality in the U.S. automobile industry", *Management Science*, Vol. 62 No. 7, pp. 1860-1877, doi: 10.1287/mnsc.2015.2195.

Hirsch, J. (2014), "Recent development in aluminium for automotive applications", *Transactions of Nonferrous Metals Society of China (English Edition)*, Vol. 24 No. 7, pp. 1995-2002, doi: 10.1016/S1003-6326(14)63305-7.

Hong, B. and Zhen, D. (2012), "An extended keyword extraction method", *Physics Procedia*, Vol. 24, pp. 1120-1127, Elsevier BV, doi: 10.1016/j.phpro.2012.02.167.

Hu, Q., Bai, Y., Zhao, J. and Cao, W. (2015), "Modeling spare parts demands forecast under two-dimensional preventive maintenance policy", *Mathematical Problems in Engineering*, Vol. 2015, pp. 1-9, doi: 10.1155/2015/728241.

Ildarkhanov, R. (2023), "Analysis of the impact of the automobile service center on the environment", *Lecture Notes in Networks and Systems*, Vol. 575 LNNS, pp. 2272-2281, doi: 10.1007/978-3-031-21219-2_255.

Jatnika, D., Bijaksana, M.A. and Suryani, A.A. (2019), "Word2vec model analysis for semantic similarities in English words", *Procedia Computer Science*, Vol. 157, pp. 160-167, doi: 10.1016/j.procs.2019.08.153.

Jerz, G. and Winterhalter, P. (2020), "The 10th international conference on countercurrent chromatography held at Technische Universität Braunschweig, Braunschweig, Germany, August 1-3, 2018", *Journal of Chromatography A*, Vol. 1617, 460698, doi: 10.1016/j.chroma.2019.460698.

Jiang, W., Zhao, W., Du, L., Zhang, K. and Yu, M. (2023), "Product perceptual similarity evaluation: from attributive error to human knowledge hierarchy", *Journal of Computing and Information Science in Engineering*, Vol. 23 No. 2, doi: 10.1115/1.4054305.

Jin, X., Zhang, S. and Liu, J. (2018), "Word semantic similarity calculation based on Word2vec", *ICCAIS 2018 - 7th International Conference on Control, Automation and Information Sciences*, IEEE, pp. 12-16, doi: 10.1109/ICCAIS.2018.8570612.

Küçüker, F.U. and Yet, B. (2022), "Reliability prediction for aircraft fleet operators: a Bayesian network model that combines supplier estimates, maintenance data and expert judgement", *Journal of the Operational Research Society*, pp. 1-12, doi: 10.1080/01605682.2022.2129486.

Kumbhar, A., Savargaonkar, M., Nalwaya, A., Bian, C. and Abouelenien, M. (2019), "Keyword extraction performance analysis", *Proceedings - 2nd International Conference on Multimedia Information Processing and Retrieval, MIPR 2019*, IEEE, pp. 550-553, doi: 10.1109/MIPR.2019.00111.

Langseth, H. and Portinale, L. (2007), "Bayesian networks in reliability", *Reliability Engineering and System Safety*, Vol. 92 No. 1, pp. 92-108, doi: 10.1016/j.ress.2005.11.037.

Lee, S.G., Ma, Y.S., Thimm, G.L. and Verstraeten, J. (2008), "Product lifecycle management in aviation maintenance, repair and overhaul", *Computers in Industry*, Vol. 59 Nos 2-3, pp. 296-303, doi: 10.1016/j.compind.2007.06.022.

Li, S., Wu, Y., Xu, Y., Hu, J. and Hu, J. (2018), "A Bayesian network based adaptability design of product structures for function evolution", *Applied Sciences (Switzerland)*, Vol. 8 No. 4, pp. 1-16, doi: 10.3390/app8040493.

Lin, Z., Zhao, Y. and Pan, E. (2018), "Development strategies for improving manufacturing quality in China's automobile industry", *Chinese Journal of Engineering Science*, Vol. 20 No. 1, p. 45, doi: 10.15302/j-sscae-2018.01.007.

Liu, P., Qiu, X. and Huang, X. (2015), "Learning context-sensitive word embeddings with neural tensor skip-gram model", *IJCAI International Joint Conference on Artificial Intelligence*, 2015-Jan, pp. 1284-1290.

Ouyang, X., Zhuang, W. and Sun, C. (2019), "Haze, health, and income: an integrated model for willingness to pay for haze mitigation in Shanghai, China", *Energy Economics*, Vol. 84, 104535, doi: 10.1016/j.eneco.2019.104535.

O'Neill, B. (2009), "Preventing passenger vehicle occupant injuries by vehicle Design- A historical perspective from IIHS", *Traffic Injury Prevention*, Vol. 10 No. 2, pp. 113-126, doi: 10.1080/15389580802486225.

Pan, T., Fan, Y., Shao, L., Chen, B., Chu, Y., He, G., Pan, Y., Wang, W. and Wu, Z. (2022), "Multiple accounting and driving factors of water resources use: a case study of Shanghai", *Journal of Environmental Management*, Vol. 313, January, doi: 10.1016/j.jenvman.2022.114929.

Pecht, M., Dasgupta, A., Barker, D. and Leonard, C.T. (1990), "The reliability physics approach to failure prediction modelling", *Quality and Reliability Engineering International*, Vol. 6 No. 4, pp. 267-273, doi: 10.1002/qre.4680060409.

Prashar, A. (2022), "Towards digitalisation of quality management: conceptual framework and case study of auto-component manufacturer", *TQM Journal*, Vol. 35 No. 2019, pp. 2436-2454, doi: 10.1108/TQM-09-2022-0289.

Rada Mihalcea, P.T. (1973), "TextRank: bringing order into texts", *Comparative Biochemistry and Physiology – Part B: Biochemistry and Molecular Biology*, Vol. 45 No. 4, doi: 10.1016/0305-0491(73)90144-2.

Ramos, J. (2003), "Using TF-IDF to determine word relevance in document queries", *Proceedings of the First Instructional Conference on Machine Learning*, Vol. 242 No. 1, pp. 29-48.

Romanowski, C.J. and Nagi, R. (2004), "A data mining approach to forming generic bills of materials in support of variant design activities", *Journal of Computing and Information Science in Engineering*, Vol. 4 No. 4, pp. 316-328, doi: 10.1115/1.1812556.

Rose, S., Engel, D., Cramer, N. and Cowley, W. (2010), "Automatic keyword extraction from individual documents", *Text Mining: Applications and Theory*, pp. 1-277.

Shah, P.K., Perez-Iratxeta, C., Bork, P. and Andrade, M.A. (2003), "Information extraction from full text scientific articles: where are the keywords? Information extractionfull text articlekeywordgene namedata miningtext mining", *BMC Bioinformatics*, Vol. 4 No. 20, pp. 1-9, doi: 10.1186/1471-2105-4-20.

Shaheen, S.A. and Niemeier, D.A. (2001), "Integrating vehicle design and human factors: minimizing elderly driving constraints", *Transportation Research Part C: Emerging Technologies*, Vol. 9 No. 3, pp. 155-174, doi: 10.1016/S0968-090X(99)00027-3.

Spasic, I., Liang, Y., Hicks, Y., Hu, H. and Liu, Y. (2022), "A data-driven approach for integrating hedonic quality and pragmatic quality in user experience modeling", *Journal of Computing and Information Science in Engineering*, Vol. 22 December, pp. 1-12, doi: 10.1115/1.4054155.

Takubo, N. and Mizuno, K. (2000), "Accident analysis of Sports Utility Vehicles: human factors from statistical analysis and case studies", *JSAE Review*, Vol. 21 No. 1, pp. 103-108, doi: 10.1016/S0389-4304(99)00057-0.

Tao, F., Qi, Q., Liu, A. and Kusiak, A. (2018), "Data-driven smart manufacturing", *Journal of Manufacturing Systems*, Vol. 48, pp. 157-169, doi: 10.1016/j.jmsy.2018.01.006.

Uusitalo, L. (2007), "Advantages and challenges of Bayesian networks in environmental modelling", *Ecological Modelling*, Vol. 203 Nos 3-4, pp. 312-318, doi: 10.1016/j.ecolmodel.2006.11.033.

Viscusi, W.K. and Evans, W.N. (2016), "Price and availability tradeoffs of automobile insurance regulation author (s): Henry Grabowski, W. Kip Viscusi and William N", Vol. 56 No. 2, pp. 275-299, Evans Published by: American Risk and Insurance Association Stable URL, (accessed 15), available at: http://www.jstor.org/stable/252989

Wagner, E., Bracquené, E. and Jaeger-Erben, M. (2021), "Exploring 14 years of repair records – information retrieval, analysis potential and data gaps to improve reparability", *Journal of Cleaner Production*, Vol. 281, 125259, doi: 10.1016/j.jclepro.2020.125259.

Wang, M., Chen, W., Fu, Y. and Yang, Y. (2015), "Analyzing and predicting heterogeneous customer preferences in China's auto market using choice modeling and network analysis", *SAE International Journal of Materials and Manufacturing*, Vol. 8 No. 3, pp. 668-677, doi: 10.4271/2015-01-0468.

Wang, S., Wan, J., Li, D. and Zhang, C. (2016), "Implementing smart factory of industrie 4.0: an outlook", *International Journal of Distributed Sensor Networks*, Vol. 2016 No. 1, p. 3159805, doi: 10.1155/2016/3159805.

Xiong, L. and Yao, Y. (2021), "Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm", *Building and Environment*, Vol. 202 June, 108026, doi: 10.1016/j.buildenv.2021.108026.

Zhang, Z., Waszink, A. and Wijngaard, J. (2000), "An instrument for measuring TQM implementation for Chinese manufacturing companies", *International Journal of Quality and Reliability Management*, Vol. 17 No. 7, pp. 730-755, doi: 10.1108/02656710010315247.

Zhang, X., Rao, R., Xie, J. and Liang, Y. (2014), "The current dilemma and future path of China's electric vehicles", *Sustainability (Switzerland)*, Vol. 6 No. 3, pp. 1567-1593, doi: 10.3390/su6031567.

Zhang, F., Liu, Z., Zhangzhong, L., Yu, J., Shi, K. and Yao, L. (2020), "Spatiotemporal distribution characteristics of reference evapotranspiration in shandong province from 1980 to 2019", *Water (Switzerland)*, Vol. 12 No. 12, pp. 1-16, doi: 10.3390/w12123495.

Zheng, J., Zhou, Y., Yu, R., Zhao, D., Lu, Z. and Zhang, P. (2019), "Survival rate of China passenger vehicles: a data-driven approach", *Energy Policy*, Vol. 129 February, pp. 587-597, doi: 10.1016/j.enpol.2019.02.037.

Corresponding author
Youyi Bi can be contacted at: youyi.bi@sjtu.edu.cn