Transportation Research Part C 90 (2018) 226–246

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Recent applications of big data analytics in railway transportation systems: A survey☆

Faeze Ghofrani[a], Qing He[a,b,*], Rob M.P. Goverde[c], Xiang Liu[d]

[a] Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York, 212 Ketter Hall, Buffalo, NY 14260, USA
[b] Department of Industrial and Systems Engineering, University at Buffalo, The State University of New York, 342 Bell Hall, Buffalo, NY 14260, USA
[c] Department of Transport and Planning, Delft University of Technology, Stevinweg 1, 2628 CN Delft, The Netherlands
[d] Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, 96 Frelinghuysen Road, Piscataway, NJ 08854, USA

ARTICLE INFO

Keywords:
Survey
Big data
Data analytics
Railway transportation systems

ABSTRACT

Big data analytics (BDA) has increasingly attracted a strong attention of analysts, researchers and practitioners in railway transportation and engineering. This urges the necessity for a review of recent research development in this field. This survey aims to provide a comprehensive review of the recent applications of big data in the context of railway engineering and transportation by a novel taxonomy framework, proposed by Mayring (2003). The survey covers three *areas* of railway transportation where BDA has been applied, namely operations, maintenance and safety. Also, the *level* of big data analytics, types of big data *models* and a variety of big data *techniques* have been reviewed and summarized. The results of this study identify the existing research gaps and thereby directions of future research in BDA in railway transportation systems.

1. Introduction

The fast-paced development of advanced technologies has made BDA as one of the most focused areas of both academia and industry. The features of big data can be characterized by 5 V, namely, volume, variety, velocity, veracity, and value (Fosso Wamba et al., 2015). The magnitude of data is featured by volume and it is among the most challenging issues specifically in terms of the storage capacity of devices (Emani et al., 2015). Variety refers to the various resources from which data can be generated in the forms of structured, semi-structured or unstructured data (Tan et al., 2015). Speed of generating data is characterized by velocity which, according to Assunção et al. (2015), may be processed in batch, real-time, nearly real-time, or streamlines. Since many data sources contain a specific level of uncertainty, the level at which a data source is trustable is featured by veracity (Gandomi and Haider, 2015). Finally, the process of revealing underexploited values from big data to support decision-making is referred by value (Assunção et al., 2015).

Railways are among the industries in which the application of big data analytics is a topic of big interest. A systematic consideration of the use of data in the context of railway transportation systems (RTS) was firstly provided in Faulkner (2002) in which four categories of data were introduced for the railway control system: (1) *configuration data* which is mostly regarded as static data that represents the entities from the real world, and change only in response to the action of maintenance or modification on these

☆ This article belongs to the Virtual Special Issue on "Big Data Railway".

* Corresponding author at: Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York, 212 Ketter Hall, Buffalo, NY 14260, USA.
E-mail address: qinghe@buffalo.edu (Q. He).

https://doi.org/10.1016/j.trc.2018.03.010
Received 11 September 2017; Received in revised form 15 January 2018; Accepted 14 March 2018
0968-090X/ © 2018 Elsevier Ltd. All rights reserved.

entities, (2) *train schedule* which is used to describe the use of the infrastructure, (3) *status data* which is provided through interfaces to external reporting systems and (4) *operational data* which accounts for the individual operational conditions which are commonly communicated to the railway control system via manual input. In this essence, many aspects of the railway world benefit from today's capability of information technology in collecting, storing, processing, analyzing and visualizing large amounts of data as well as new methods coming from machine learning, artificial intelligence, and computational intelligence to recognize patterns and retrieve useful information (Al-Jarrah et al., 2015). This is in accordance with the growing demand on railway transportation, which necessitates ensuring customer satisfaction by being safe, reliable and service-oriented. In fact, railway industry has been revolutionized by big data analytics (BDA) which contributes to the decision-making processes of railway companies. Studies in the literature demonstrate multiple advantages of applying BDA in RTS in reducing cost and delay, and in parallel maintaining high standards of safety, reliability and customer satisfaction.

Despite the fact that BDA adoption can enhance RTS performance, not many of railway-related enterprises have implemented BDA in one or more RTS areas. This is mainly due to the lack of understanding on how BDA can be implemented in RTS, the inability to collect and process massive data, data security issues, routinization and assimilation of BDA by railway companies. This motivates our exploration of identifying the existing gaps in the applications of BDA in RTS.

There are a number of surveys in the literature on the application and challenges of BDA in RTS context. However, most of these studies tend to focus on a specific aspect of RTS. For instance, in Hodge et al. (2015), a survey of wireless sensors network technology for monitoring and analyzing railway systems, structures, vehicles, and machinery was conducted. A survey of railway-related planning and scheduling issues in Europe was provided in Turner et al. (2016). As another example Nunez and Attoh-Okine (2015) conducted a literature review on the application of metaheuristic optimization in railway engineering. Some other survey articles on the application of data analytics in a specific aspect of RTS can be found in Soleimanmeigouni, et al. (2016), Singh et al. (2015), Hodge et al. (2015), Thaduri et al. (2015), Griffin et al. (2014), Summit (2014), Figueres-Esteban et al. (2015). To the best of authors' knowledge, the literature in this field of study suffers from the lack of a holistic survey which takes a broad perspective of RTS as a whole and cross-maps with BDA.

Our survey develops a taxonomy framework in Section 2, which identifies the areas of RTS and connects them with the level of analytics, BDA models, and techniques. The developed framework aims to provide a complete picture of where and how BDA has been applied in RTS. To obtain this objective the study considers four aspects namely, the *areas* of railway transportation in which big data analytics is applied, the *level* of big data analytics in rail transportation, types of big data *models* and big data *techniques* used to apply these models.

Section 3 specifically studies the material evaluation from the proposed framework. This includes the review of articles based on the three areas of railway transportation systems, i.e. maintenance, operations, and safety, as well as review by BDA models and BDA techniques. It should be mentioned that although the term "operations" is usually referred to a comprehensive spectrum of activities in the RTS context which occasionally include maintenance and safety as well, what we mean in this paper is the activities related to the train traffic and transportation services, thus excluding maintenance and safety activities.

One of the limitations of the current paper is that the categories in the proposed classification framework are interpretative, which is probably to result in subjective bias. This is also one of the main issues of the content analysis method according to Seuring (2013), despite several of validations being carried out.

In Section 4 a discussion on the future direction of the studies of BDA in the context of RTS as well as the advanced big data computational technologies in railway transportation systems around the world is provided and finally the conclusions are provided in Section 5.

2. Methodology

According to Brewerton and Millward (2001), from a methodological point of view, a literature review could be as comprehended as a content analysis which considers both quantitative and qualitative aspects of a context. In this essence, we have considered a survey approach based on content analysis by Mayring (2003). Practices and context of content analysis by Mayring (2003) has been taken up by various scholarly communities and modified by some of them specifically in the field of supply chain management (Seuring and Müller, 2008; Brandenburg et al., 2014; Govindan et al., 2015; Seuring, 2013; Klewitz and Hansen, 2014). With respect to BDA, this methodology has been adopted for big data analytics in supply chain management by Nguyen et al. (2017). Apart from the field of supply chain management, Gläser and Laudel (2010) took Mayring's method as a starting point for developing their own technique for analyzing expert interviews in social science. All the contexts which apply content analysis involve systematic reading or observation of texts, documents or artifacts which are assigned labels to indicate the presence of interesting, meaningful patterns in a specific field of study (Tipaldo, 2014). After labeling a large set of texts, documents, papers (in our case) researcher is able to statistically estimate the proportions of patterns in those documents, as well as correlations between them. This refers to the fact that content analysis is not limited to specific contexts to be applicable. As long as an eligible topic for a literature review is defined, sufficient materials are observed for a topic and the questions on "How the data is collected?", "Which population the data is taken from?", "How the relevant context is defined?", "What are the boundaries of the analysis?" are responded, the content analysis technique could be conducted as a methodology for classification and labeling of available materials (Krippendorff, 2004). It should be noted that in content analysis the analyst can make various decisions about how the paper is to be comprehended and what dimensions/classification are about to be considered. This is such a risk which can be reduced by involving two or more researchers when searching for and analyzing the data (Krippendorff, 2004) which is true for our research. Our survey is conducted according to the four-step iterative process as follows (Mayring, 2003; Seuring and Müller, 2008):

a. Material collection which defines the material to be collected as well as the unit of analysis (i.e., the single paper).
b. Descriptive analysis which describes the formal and general aspects of the studied topic.
c. Category selection in which structural dimensions form the major topics of analysis are considered to provide analytic categories of the collected material.
d. Material evaluation which aims to analyze the material according to the proposed framework and identify the relevant issues and interprets the results.

It might be necessary to redo steps *c* and *d*, as the dimensions and underlying categories need to be revised (Mayring, 1990). We will discuss the first three steps in the following subsections. However, the last step is discussed in greater details in Section 3.

2.1. Material collection

Research delamination by defining boundaries is among the most important steps to conduct a literature review (Seuring and Müller, 2008). To this end, we have considered the following specific features as criteria for our survey:

1. Before searching for articles, we have identified data-related keywords such as "Data analytics, Big data, Data mining, Machine learning, Descriptive analytics, Predictive analytics" to use together with RTS-related keywords including "Rail, Railway Engineering, Railway Systems, Railway Operations, Railway Safety, Railway Maintenance" as effective sets of keywords to capture the synthesis of existing literature related to our research topic.
2. This analysis aimed only at papers in scientific journals, conferences and dissertations in English for the last 15 years, from 2003 to 2017, with a data analysis focus in RTS.
3. Publications about qualitative challenges of BDA in RTS or the ones with general introduction on the areas of RTS where BDA will play a role were not considered. This review only includes the studies with quantitative results as well as surveys on the application of data analytics in RTS.
4. Articles with pure mathematical modeling of RTS problems are not included. The modeling should somehow have been applied on a data set to be considered for our survey. The dataset might be real and with considerable size such as the ones received by detectors and sensors or small as the ones obtained by simulation or field tests.

We have used major databases to search for related articles, such as those provided by major publishers including Science Direct, Emeralds, Scopus, EBSCO, and IEEE Xplore. Cited references of studied papers were also used as a source for finding the related articles. Taking the mentioned considerations and boundaries into account, a total of 115 papers were identified.

2.2. Descriptive analysis

In Fig. 1, the distribution of published papers in this specific field of study is shown from 2003 to 2017. The notion of data analytics started from 2000 by the widespread use of computers and automatic systems in different industries around the world. In RTS context, we have focused our research for the last fifteen years during which the majority of BDA studies are presented. As expected, the number of publications has increased steadily specifically in the last ten years (apart from a spike in 2010) which implies that the application of BDA on the RTS area, is a fast-growing research field. It is worth mentioning that the 2017 articles are the ones published by September 2017 while for the other years we have the full year data. This is probably the main reason for the lower number of publications in this year compared to the other years.

In Fig. 2, the distribution of reviewed articles based on the type of publication is shown. According to this figure, the majority of

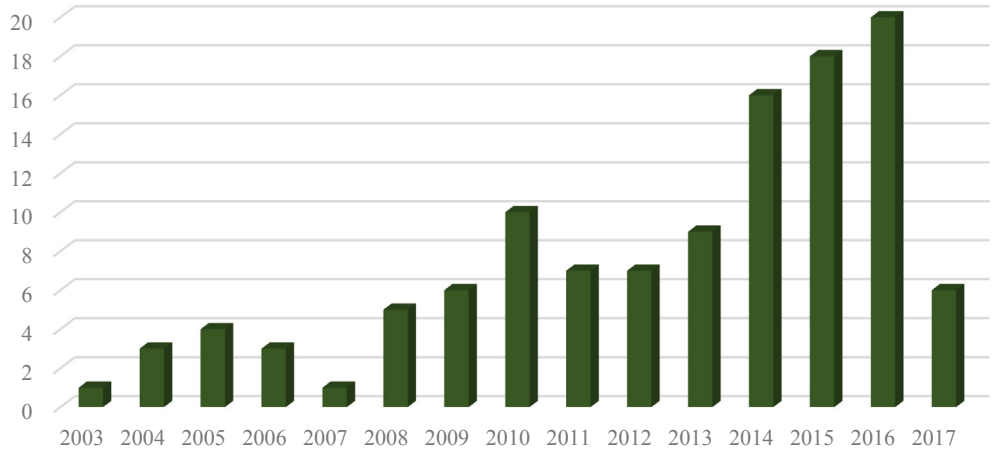

Fig. 1. Distribution of reviewed articles by year.

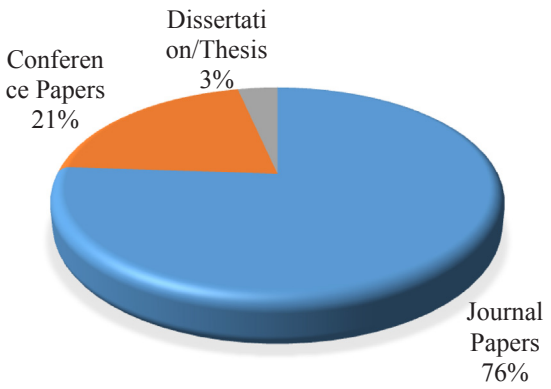

Fig. 2. Distribution of reviewed articles by type of publication.

articles are the journal articles (89 out of 115 articles) out of which 9 articles are survey papers in this field of study and the other 106 papers are about the application of BDA in RTS. Conference papers and dissertations account for 21% and 3% of articles, respectively.

It is worth mentioning that one might think that some of the most recent developments in some RTS areas might appear in magazines of rail industry rather than scientific papers. To name a few, we can refer to Global Railway Review (2017), Progressive Railroading (2017) and Tramways and Urban Transits (2014). These magazines are all among famous magazines which consider the discussion on the application of BDA in the context of RTS. However, to avoid conflict of commercial interest, and for the sake of scientific rigor, we only build our focus on peer-reviewed academic articles. Most of the studies we have reviewed are from reliable scientific journals, conferences and other reliable research resources which is a guarantee for the validation of our study.

We utilize the size and type of data as useful properties to classify the articles. Small data is usually collected to answer the problem at hand. For small data, there is control of the data. Once the data is collected, it is ready for analysis. However, big data involves multiple datasets and a complicated structure. This is the main reason for which the data takes a long time to be cleaned and processed. The data analyst has to come up with relationships in the data structures. Then different algorithms are used to verify the findings. For the purpose of the current study, the data of articles in which multiple datasets from different sources are merged to establish the main dataset and the data-cleaning and preprocessing account for a huge proportion of computation have been considered as “big”. Moreover, we have considered the data sets with over 50k variables or 50k observations as big. As one can see in Fig. 3, the data used in the majority of articles in this field of study has been big and real. Also, it is found from this figure that the majority of articles in which simulation/test data is analyzed, the data size is small.

The popularity of small and big data analytics over the years of study is shown in Fig. 4. As expected, before 2007, the application of small data analytics dominated the literature while the opposite story is true after 2007. This could be referred to the reason that the term “Big Data” emerged back in 2007 (Nguyen et al., 2017) after which it started to become a widespread phenomenon to hit a peak in 2010.

2.3. Category selection

As mentioned before, category selection provides us structural dimensions that form the major topics of analysis. In order to address the main dimensions of our research, we have selected a four-layer structure provided in Fig. 5. Each layer in this structure represents the key topics for each of the four dimensions in our study. The first layer represents the main areas of RTS: Maintenance, Operations, and Safety. In the second layer, the BDA-RTS literature has been divided into three categories of descriptive, predictive and prescriptive analytics. This is a common classification widely used in BDA studies (Delen and Demirkan, 2013; Duan and Xiong, 2015). The simplest level of analytics is descriptive analysis which focuses on the past events whereas predictive analytics

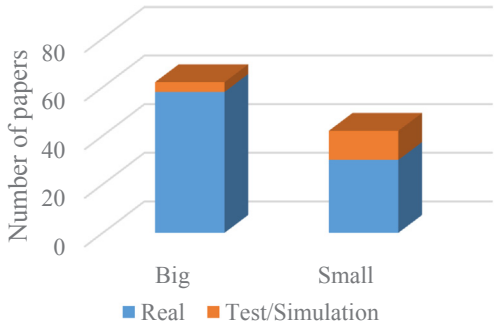

Fig. 3. Distribution of reviewed articles by type and size of data.

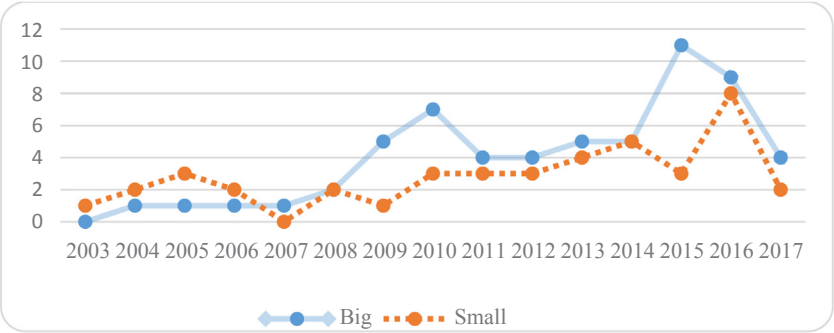

Fig. 4. Distribution of reviewed articles by size and year.

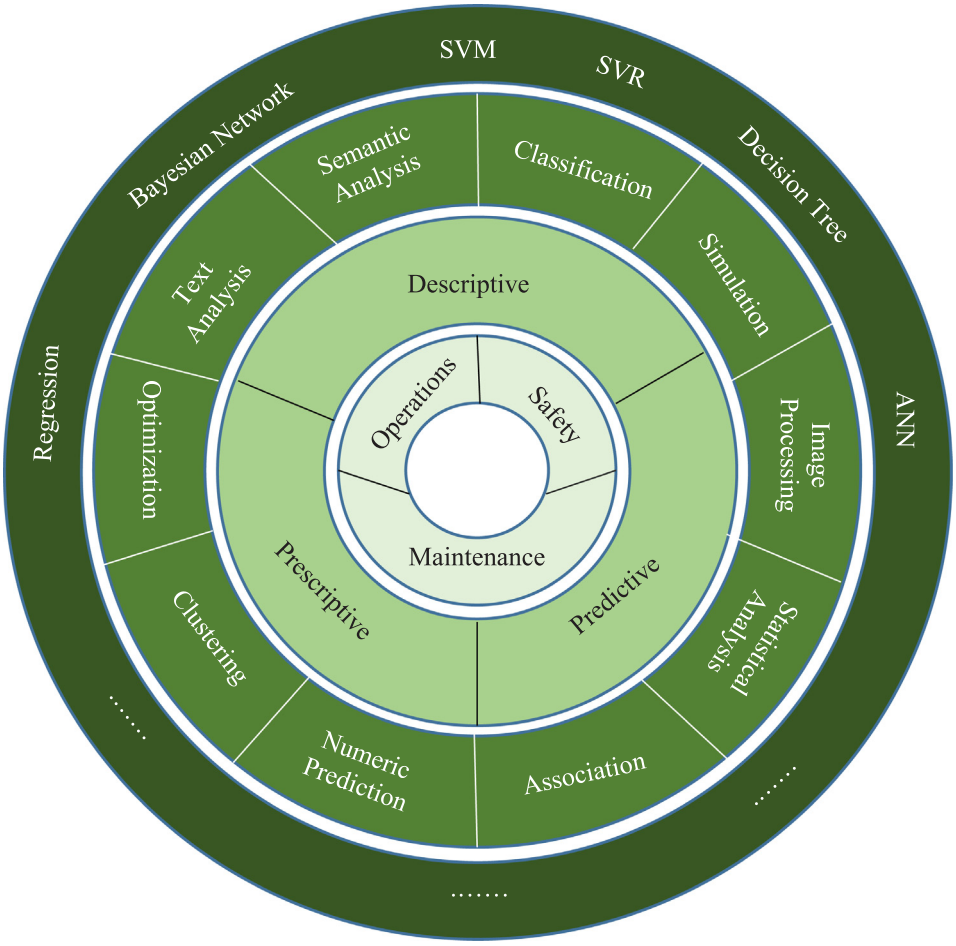

Fig. 5. Four-Layer Structure for reviewed papers.

concentrates on future events and prescriptive analytics is dedicated for decision making (Rehman et al., 2016). The third layer contains the most widely used models of BDA (Erl et al., 2016), such as clustering, classification, association, simulation, etc. Finally, the final layer includes the popular techniques used to implement BDA models. The explanation on these layers as well as their corresponding articles would be discussed in details in the following section.

3. Material evaluation

3.1. Classification based on RTS areas

As mentioned earlier, there are three main RTS areas-maintenance, operations and safety-which have been benefitted by BDA

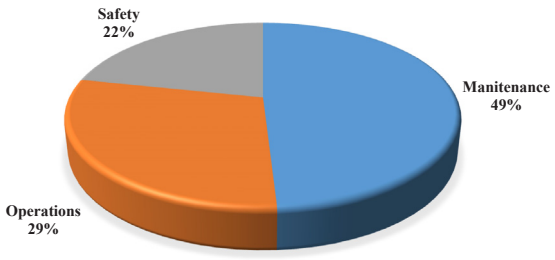

Fig. 6. Distribution of articles by RTS areas.

studies in the recent years. The distribution of BDA studies in each of these RTS areas is shown in Fig. 6. According to this figure, maintenance studies have been the most popular RTS area studied by researchers as it accounts for almost half of the reviewed papers (52 out of 106 articles excluding survey papers).

3.1.1. Maintenance

Maintenance of RTS can be categorized either based on the type of the maintenance (corrective, preventive and condition-based) or based on the system component (Vehicle, Track, Signaling Equipment) which receives the maintenance. These two categories will be discussed in the following subsections.

3.1.1.1. Classification based on the type of maintenance: Corrective, preventive and condition-based. A maintenance activity is usually defined as a set of activities performed to maintain the functionality of an item or system (Budai-Balke, 2009). Most studies divide the maintenance of deteriorating systems including transportation systems into three main categories: Corrective Maintenance (CM), Preventive Maintenance (PM) and Condition-Based Maintenance (CBM) (Budai et al., 2006; Jardine et al., 2006; Figueroa-García et al., 2015; Budai-Balke, 2009). Corrective maintenance (CM) which is also known as reactive maintenance is a kind of strategy undertaken after a defect or failure occurs. This strategy leads to high levels of maintenance costs due to sudden failure and system recovery (Figueroa-García et al., 2015). Different than CM strategy, preventive maintenance (PM) strategy involves the performance of maintenance activities before the failure of equipment. PM includes scheduled adjustments, major overhauls, replacements, renewals, and inspections. PM can be carried out either on the system downtime or while the system is in operation. The most significant advantage of PM is that it can be planned in advance and performed when convenient (Budai-Balke, 2009).

In CBM, the main objective is to optimize maintenance activities given the estimation of the component actual status using monitoring and inspection techniques. This results in discovering those components where maintenance is required so that the maintenance cost is greatly reduced. Further, predictive maintenance leverages the prediction of the failure time to proactively schedule maintenance activities. Actually, we can regard predictive maintenance as a part of CBM, since CBM involves more or less certain type of prediction.

Prognosis of Remaining Useful Life (RUL) is among the most studied problems in predictive maintenance (Li and He, 2015). Providing a higher level of safety and reducing maintenance costs, preventive and conditional maintenance tasks, nowadays, have attracted more interest compared to corrective maintenance (Fumeo et al., 2015).

The distribution of railway maintenance types in the reviewed BDA articles is presented in Fig. 7. As observed, preventive maintenance receives the highest level of focus in the recent research on railway maintenance, while corrective maintenance is less considered in the literature.

3.1.1.2. Classification based on RTS components: vehicle, track, and signaling equipment. The past studies on RTS maintenance were carried out more specifically on railway vehicles, track or signaling equipment. Fig. 8 depicts the proportion of studies in the literature for each type of these components. What really matters for each of these components is the procedure in which a failure event is identified.

The analysis of condition data of rail vehicle contributes to the vehicle set-outs and maintenance planning. According to Association of American Railroad, monitoring condition of railway vehicles is mainly performed using acoustic bearing detectors, hot

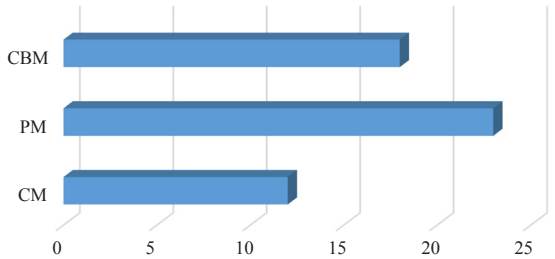

Fig. 7. Distribution of maintenance-related articles by type of maintenance.

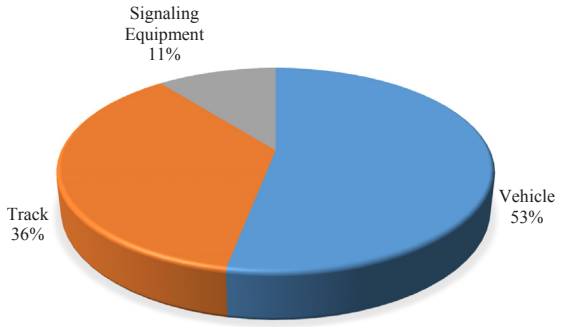

Fig. 8. Distribution of RTS components for maintenance.

box detector (HBD) temperature trending, hot/cold wheel detectors, truck performance detectors, hunting detectors, wheel impact load detectors (WILD), cracked axle detectors, cracked wheel detector and machine vision. In most cases, defective wheels generate high impact load on the track which is detected by WILD as it weighs each wheel several times when the wheel passes by a detector in a certain distance (Li et al., 2014). Strain-gauge-based technologies are used by WILD to measure the performance of a railcar in a dynamic mode by quantifying the force applied to the rail (Stratman et al., 2007). Once a train is detected, WILD generates different levels of data including train data, equipment data, truck data and wheel data as shown in Table 1 (Wang et al., 2017).

With respect to the track maintenance, an optimal maintenance decision relies on massive dynamic and static datasets from different sources, including service failure data, signal data, ballast history, grinding history, remedial action history, traffic data, inspection data, as well as curve and grade data as shown in Fig. 9.

The track inspection can be mainly represented in two different ways: track geometry inspection or track structure inspection. There are numerous studies in the literature on modeling the geometric and structural degradation of railway track which is a prerequisite for track maintenance planning. Track geometry degradation refers to severe ill-condition in geometry parameters such as profile, alignment and gauge, while the structural condition of track includes the condition of rail, ballast, ties system, sub-grade and drainage system. Presence of structural defects such as cracks and geometry defects like track misalignments is a major threat to the safe operation of a railway system. A complete list of geometry defects can be found in He et al. (2015) and Zarembski (2015). Rail defects occur due to wear (primarily in curves), fatigue (in the form of surface/subsurface initiated cracks), and plastic flow (in the form of Corrugation in rails). There are several types of rail defects. Some of defects are suitable to be modeled by data-driven models while the other ones are suitable for mechanistic models. For more details the types of rail defects readers can refer to One and Pérez (2003). During the process of the rail operations, defects can worsen if no recovery action is undertaken. They may finally develop to complete rail breakage, which is a major cause of train derailment. Therefore, track inspection cars are used to detect the defects before they develop to complete rail breakage. Two common types of monitoring cars are Ordinary Measurement cars which measure the rail geometry and surface deterioration and Ultrasonic Inspection (USI) cars which measure rail breakage and internal cracks (Podofillini et al., 2006). Drones are also among those tools that have gained popularity during the recent years for track inspections. Images are usually processed from the front camera of the drones (Pall et al., 2014). It is expected that fast image processing and analysis will be emerging in drone-base track inspection.

Track degradation models can be either based on the physical laws describing the behavior of the asset known as mechanistic models or based on data-driven models which mainly relies on machine learning algorithms (Fumeo et al., 2015). The focus of this study has been mainly on data-driven models, however some examples of mechanistic models for railway maintenance in the reviewed articles could be found in Liu et al. (2006), Morgado et al. (2008) and Sura (2011). A comprehensive review of mechanistic models is also presented in Singh et al. (2015).

According to Fig. 8, a few of the studies on RTS maintenance focuses on data analysis of signaling equipment including turnouts,

Table 1
WILD Data and data connections (Wang et al., 2017).

Source of data	Explanation	Corresponding attributes
Train	Each train has at least one locomotive and several equipment. That equipment may belong to other companies	Unique identifier for an e-detector, Count of rail cars in a train, Max peak wheel load reading in kips, Max ratio between average& kips, Locomotive name, The max train car truck hunting index, Kind of train
Equipment	Each equipment has at least two trucks. Or more than 20 cars in rare cases	Unique identifier for an e-detector, Equipment initial, Whether the axle count for a car was correct, Truck amount in one equipment, User id
Truck	Each truck has two axles. These trucks could be different types	Unique identifier for an e-detector, The sequence of a truck on a car, The weight of a truck recorded in tons, The truck hunting index for each truck, Timestamp indicator
Wheel	Each truck has two axles. These trucks could be different types	Unique identifier for an e-detector, Identifier of an axle on specific wheel, Average load reading in kips for a wheel, Peak load reading kips for a wheel, Ave lateral load reading kips for a wheel, Peak lateral load reading kips for a wheel

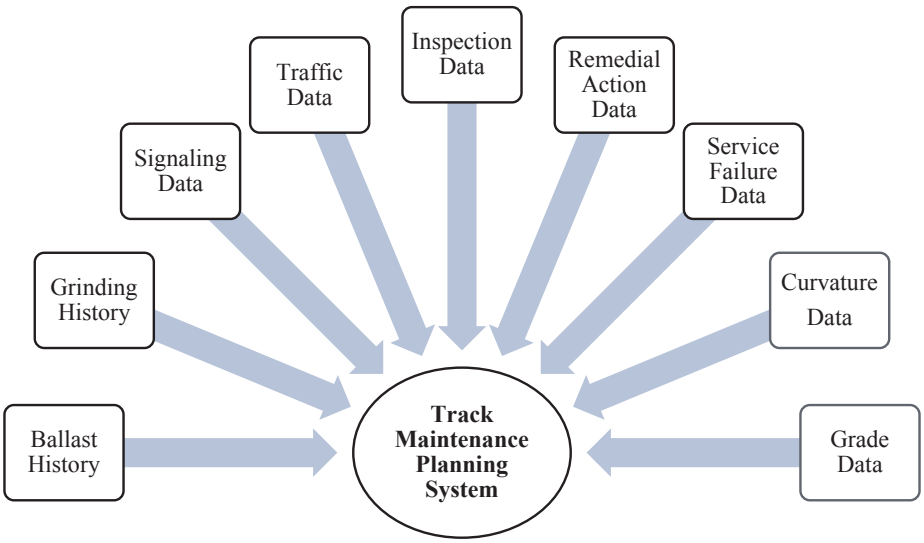

Fig. 9. Data required for track maintenance planning system.

track circuits, etc. Signaling systems, like most electronic equipment, must have hardware maintenance; when parts are replaced, the software may be updated or not. This depends mainly on whether the new electronic device is configured by the factory or must be adapted to the particular specifications of its location and functions (Morant et al., 2012).

The classification of studied articles in maintenance is presented in Table 2.

According to Table 2, it is observed that:

- CM has been mainly studied for tracks but not for vehicle and signaling equipment.
- CBM has been the most studied type of maintenance for vehicle while PM has been the most popular one for track maintenance.
- Few studies have considered the PM on other vehicle components rather than wheels.
- CBM for locomotive and traction has not been studied in the reviewed articles.
- The CM has been studied mainly for the track structure deterioration models compared to track geometry models.

3.1.2. Operations

Intelligent Rail Transportation Systems (IRTS) have provided innovative technologies for railway infrastructure managers and train operating companies that help them to make more efficient decisions. The application of BDA in RTS operations ranges from train delay analysis and prediction to passenger route choice and demand forecasting with the computerized processing of massive amounts of train positioning and passenger travel data. These big data analytics improve timetabling and simulation models and allow improved decision making in real-time rail traffic management. Table 3 gives an overview of the big data sources of railway operations, which will be considered in this section.

The automatic signaling systems of today are based on track-clear detection and include train describer systems that keep track of train movements by their train description (train identification number), which is the basis for automatic route setting and centralized traffic control. Train describer systems log the incoming interlocking and track occupation messages together with the train describer generated messages, but these logs were only kept for a few days to support the investigation of possible accidents (Goverde and Hansen, 2000). Only at the start of the 21st century, the logs became archived when railway companies realized these data could be used for railway operations analysis. Collecting, processing and transforming these train describer record data in combination with timetable data for descriptive analysis of train delays and timetable improvements can be seen as the first applications of big data in railway operations. Process mining with domain knowledge was used to process train describer records and retrieve train positions on track occupation level (Goverde and Hansen, 2000; Daamen et al., 2009; Kecman and Goverde, 2012), as well as route conflicts from which secondary delays due to unscheduled braking and waiting in rear of signals could be derived (Daamen et al., 2009; Goverde et al., 2008). This information was also used to automatically derive conflict trees of successive route conflicts with associated secondary delays to identify the impact of timetable shortcomings and root causes of secondary delayed trains (Goverde and Meng, 2011); The train describer systems are also used to derive train delays at stations online, which is another source of large data used originally for punctuality statistics, although this delay data is less accurate. More recent, also alternative train positioning data were used for data analysis, including train event recorder data (De Fabris et al., 2008) and GPS data (Medeossi et al., 2011).

The train describer data enabled data-driven predictive models using historical and real-time data, with methods including robust regression, regression trees, and random forests (Kecman and Goverde, 2015a), and event graph models for online track conflict and train delay predictions (Kecman and Goverde, 2015a; Hansen et al., 2010). Delay data has been used for deriving dependencies between trains using data mining (Flier et al., 2009) and decision trees (Lee et al., 2016), for stochastic delay prediction over large networks using event graph models (Berger et al., 2011), for train delay prediction using regression models (Wang and Work, 2015)

Table 2
Summary of literature by RTS maintenance.

RTS component		CM	PM	CBM
Vehicle	Wheel	Grassie (2005)	Braghin et al. (2006), Sura (2011), Hajibabai et al. (2012), Papaelias et al. (2016), Skarlatos et al. (2004), Yang and Letourneau (2005) and Yang and Létourneau (2009)	Bladon et al. (2004), Enblom and Berg (2005), Adam et al. (2013), Palo (2014), Li and He (2015), Li et al. (2014), Starke et al. (2006), Stratman et al. (2007), Schlake et al. (2010) and Bevan et al. (2013)
	Truck	_	Shafiullah et al. (2010)	Bladon et al. (2004), Palo (2014) and Li and He (2015)
	Bearing	_	Papaelias et al. (2016)	Enblom and Berg (2005), Loutas et al. (2013), Fumeo et al. (2015) and Vale et al. (2016)
	Traction	_	Sammouri et al. (2013) and Morgado et al. (2008)	_
Track	Geometry	He et al. (2015) and Nunez et al. (2015)	Budai et al. (2009), Budai-Balke (2009), Hu and Liu (2016), Zhu et al. (2013), Sadeghi and Askarinejad (2012), Zarembski et al. (2016) and Shang and Berenguer (2014)	Giben et al. (2015) and Sharma et al. (2017)
	Structure	Tyler Dick et al. (2003), Liu et al. (2006), Schafer and Barkan (2008), Corbetta et al. (2015) and Jamshidi et al. (2017)	Kumar (2006), Budai et al. (2009), Zhi et al. (2016), Chattopadhyay and Kumar (2009), Sadeghi and Askarinejad (2012) and Zarembski et al. (2016)	Kaewunruen (2014) and Su et al. (2016)
Signaling equipment		Zhao et al. (2014)	Zhao et al. (2013), Zhang et al. (2016) and Sammouri et al. (2013)	Yin and Zhao (2016), Kaewunruen (2014) and Yilboga et al. (2010)

Table 3
Big data sources of railway operations.

Big data sources	Typical contents
Train describer data	Track occupation and release times, train description steps, signal states (stop/go), switch states (left/right)
Traffic control delay data	Delays at stations or other timetable points
GPS data	Train positions
Train event recorder data	Train positions and speeds, traction, brake applications
Traffic control incident registration data	Begin and end time of disruptions, failing elements
Timetable data	Arrival and departure times, train routes, stops
Ticket sales data	Tickets available
Automatic Fare Collection data (smart card data)	Passenger check-in and check-out times
Website data	Timetables, recommended travels and prices, train delays, disruption locations and times, online ticket sales

and support vector regression (Markovic et al., 2015), and with weather data for delay prediction over large networks using extreme learning machines (Oneto et al., 2016, 2017a, 2017b) and kernel estimates and random forests (Oneto et al., 2016). Delay data from simulations have been used to predict delays based on train mix, operating parameters and network topology using regression models (Murali et al., 2010). Another application of train position data has been in estimating train characteristics using graphical tools for train event recorder data (De Fabris et al., 2008), Simulated Annealing for GPS data (Medeossi et al., 2011), and Genetic Algorithm for track occupation data (Bešinović et al., 2013). GPS train positioning data reduction is a large scale combinatorial problem which is considered in Chen et al. (2010).

Ticket sales data has been used for short-term rail demand forecasting using Artificial Neural Networks (ANN) (Tsai et al., 2009) and hybrid Support Vector Machines (SVM) (Jiang et al., 2014; Sun et al., 2015). In combination with delay data, econometric regression models have been developed to estimate the impact of lateness on demand (Batley et al., 2011). Open data from ticket websites is a recent source of big data, including the automatic collection of 'remaining ticket data' to derive train demand (Wei et al., 2017).

Automatic fare collection systems using contactless smart cards to check in and out of stations and/or trains is a recent technology that generates big smart card transactions data. Note that destinations are not known beforehand as with ticket sales. Combined with timetable data this smart card data has been used for deducing passenger route (or train) choice. Methods include path searching over event networks (Kusakabe et al., 2010; Van Der Hurk et al., 2015), regression (Sun et al., 2012), and Markov Chain Monte Carlo (MCMC) methods (Sun et al., 2015). The route choice models can be used to simulate the train load factors of new timetables and their rolling stock assignments (Jiang et al., 2016), but they are also useful in the case of disruptions when traditional route choice models are not valid (Van Der Hurk et al., 2015). Moreover, these models can be used to derive passenger punctuality instead of train punctuality and in decision making during disruption management.

Rail traffic management can profit from big data analytics by estimating the impact of traffic control decisions on train delays, passenger route choice and resulting train loads, and passenger delays. In the case of disruptions, disruption lengths can be estimated using Bayesian networks based on details about the disruptions (Zilko et al., 2016). However, data about disruptions are very poor and often do not include much more than the failing element (rolling stock, track circuit, signal, point, etc.) in free-form text fields of rail traffic control information systems, while information on the underlying failing component and required repair are not available. Collecting this kind of data and making them available is essential for the application of big data analytics to improve decision making during disruptions.

Another source of big railway operations data is the World Wide Web using web scraping of open data that is provided on websites such as timetables, travel recommendations and prices, dynamic departure times, and disruptions and engineering works. The automatic collection of 'remaining ticket data' to derive train demand is one example of this (Wei et al., 2017). Other publications using these open railway operations data have not been found, but this is a promising area as more and more train operating companies increasingly share dynamic data on their websites.

3.1.3. Safety

Safety is the top priority for rail transportation systems. Although railway is currently the safest mode of surface transportation, accidents still occur. In particular, if the derailed trains transport hazardous materials or passengers, the consequences could be disastrous. Learning from historical accident data is important for understanding and preventing train accidents. Based on statistical analyses of historical data, we can gain an understanding of the cause (Evans, 2010; Liu et al., 2012, 2011; Lin and Saat, 2014), frequency (Evans, 2010; Liu, 2017; Liu et al., 2017), severity (Evans and Verlander, 1996; Liu et al., 2013; Ghomi et al., 2016) and contributing safety factors related to infrastructure (Liu et al., 2017), operations (Liu, 2016a; Liu et al., 2017) or environment (Liu, 2016b). In general, rail accident database is (fortunately) not big. The sample size of accident data can range from a couple of hundreds to thousands, depending on the time period and region of interest for the analysis. To handle this moderate size of data, summative statistics, regression models and data analytics methods have been used to understand the high-level relationship between accident risk and its influencing risk factors (e.g. Liu et al., 2011, 2017, 2012; Liu, 2016a; Mirabadi and Sharifian, 2010; Shao et al., 2016; Baysari et al., 2008).

The analysis of historical accident data provides useful, high-level views regarding safety trends and characteristics. However, it is

not always sufficient to depict and predict the localized risk profile for a specific location given a time period. To address this challenge, researchers use the "first-principal" strategy to trace the occurrence of a train accident to a series of precursor events (Kyriakidis et al., 2012). Each precursor event typically occurs more frequently and thus has a large dataset for the analysis. For example, the releases of railroad tank cars in train accidents have long been a key safety concern in the North American rail industry. This type of accident occurs rarely but can lead to disastrous consequences. The analysis of this type of accident using only historical data could lead to statistical errors, such as regression to the mean (Liu, 2015). To address the "small data" problem, researchers analyzed the precursors of a hazardous material release incident, such as train derailment, the number of tank cars derailed and the releases of derailed tank cars (Liu et al., 2014). Then, the probability of a release incident can be estimated based on the probability of these precursors using probabilistic models (Bagheri et al., 2014). Besides infrastructure and equipment failures, human error is another major accident cause (Baysari et al., 2008). In order to further reduce human errors, the US Federal Railroad Administration sponsored a voluntary confidential program called "C3RS", allowing railroad carriers and their employees to report close calls. The program provides a safe environment for employees to report unsafe events and conditions. Employees receive protection from discipline and FRA enforcement. This program has been reported to be successful in terms of reducing accident occurrence. Similar programs were also implemented in other nations (Hughes et al., 2015).

In the RTS, each accident precursor can be affected by a large number of causal or affecting factors. Some of these factors can dynamically change with time and rail operations. In order to handle complex network-level safety analysis accounting for many factors and precursors, researchers have developed system-based approaches (e.g. Bayesian network, Petri network) to estimate the risk of a rare transportation incident (Bearfield et al., 2013; Castillo et al., 2016). While the methodological framework is sound, estimating the parameters in the model and implementing the model in railroad practice is a very challenging task due to data limitations and the complexity of railway safety problem.

As stated in previous sections, rail infrastructure monitoring and train operations all involve big data sources. How to leverage these big data sources for safety analysis requires a deep understanding of their relationships with safety (Bearfield et al., 2013). At present, the clear linkage between small accident data and other big data sources is not well understood. Understanding this would not only demand advanced analytical approaches but a practical understanding of various engineering, operational and management aspects of RTS. It is worth mentioning that in most cases safety-relevant data is distributed in various different servers in the railway system. Therefore, it is less likely that all safety-relevant information is collected with a single computer cluster. In the future, more advanced technologies could be used to collect and integrate distributed data sources for systematic safety analysis (Van Gulijk et al., 2015).

3.2. Classification based on level of analytics

The aim of this taxonomy is to determine the level of analytics used to support decision-making process for each of the RTS areas. Fig. 10 depicts the distribution of level of analytics in each year. From this figure, we can see that almost through all the years of study, the application of data analysis in RTS is mainly at descriptive level (39 out of 106 articles) and predictive analysis (36 out of 106 articles) and prescriptive analysis is the least studied.

Table 4 presents the level of analytics at which each area of RTS has been studied. According to this table most of the safety-related articles are at descriptive level, while for maintenance and operations-related articles, predictive analytics dominate. It is obvious that both predictive and prescriptive studies in safety are under-studied.

3.3. Classification based on BDA models and techniques

There are several approaches in the literature that researchers have applied for analysis of data in different problems of RTS. A summary of the number of articles which used any kind of BDA models based on RTS areas is provided in Table 5. Each BDA model includes a bunch of techniques that are used in the literature. We discuss some of the most prevalent ones observed in reviewed

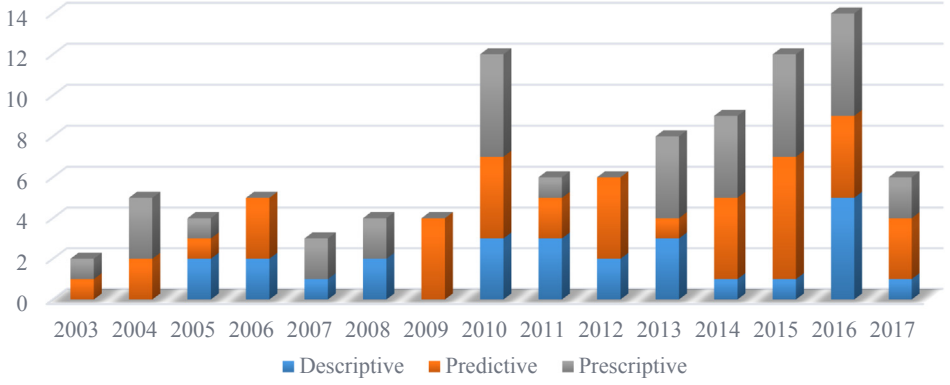

Fig. 10. Level of analytics in the studied articles by year.

Table 4
Distribution of level of analytics by RTS areas.

RTS areas/level of analytics	Descriptive	Predictive	Prescriptive	Total number of articles
Maintenance	10	22	20	52
Operations	10	14	7	31
Safety	19	0	4	23
Total number of articles	39	36	31	106

Table 5
Distribution of BDA models by RTS areas.

BDA models		Maintenance	Operations	Safety	Total
Association		5	0	4	9
Clustering		0	0	2	2
Prediction models	Classification	4	2	4	10
	Pattern Recognition	9	6	0	15
	Time Series	5	2	0	7
	Stochastic Models	1	3	0	4
Optimization-based models	Exact methods	3	1	0	4
	Heuristic methods	2	5	0	7
Text-analysis		3	0	2	5
Statistical analysis/modeling		17	9	13	39
Simulation		10	4	0	14
Visualization and image processing		3	1	0	4
Mechanistic analysis		6	0	0	6
Process mining		0	2	0	2
	Total	68	35	25	128

papers. It should be noted that some articles used more than one BDA model.

Association models, which are usually categorized at descriptive level of BDA, aim at discovering patterns of co-occurrences and strong relationships among items in a large database. Association rule mining is the most applied technique in association models (Mirabadi and Sharifian, 2010; Sammouri et al., 2013; Ghomi et al., 2016). As an example, the study by Mirabadi and Sharifian (2010) identified the hidden relationships of the most common accidents and their potential causes such as human factors, rolling stock, tracks and signaling systems.

Clustering models refer to the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. The K-means algorithm is among the most adaptable clustering techniques used in the literature (Shao et al., 2016). It is usually performed as an initial data analytics process to partition the heterogeneous data into more homogeneous segments. Clustering can be conducted also in the context of text mining by identifying key tokens within the source text to identify the theme of a record (Hughes et al., 2015).

According to Table 5, prevailing prediction models, including classification, pattern recognition models, time series analysis, and stochastic models, are well applied in the studied papers (together around 36% of the articles). These models are usually considered at the predictive level of analytics. Classification models aim at classifying data objects in a dataset into predetermined class labels. For classification models, decision tree is one popular technique used in the literature (Yang and Létourneau, 2009; Yin and Zhaos, 2016; Ghomi et al., 2016). As an example, Ghomi et al. (2016) used a decision tree to determine the class label of injury-severity of accidents at highway-railroad grade crossings considering different factors.

Pattern recognition models usually focus on the recognition of patterns and regularities in data by using more advanced machine learning techniques, including ANN (Yu et al., 2007; Yilboga et al., 2010; Chen and Gao, 2012), SVM (Hu and Liu, 2016) and Support Vector Regression (SVR) (Loutas et al., 2013; Fumeo et al., 2015; Markovic et al., 2015). Pattern recognition is the most implemented model of prediction in the reviewed papers.

Time series is another popular prediction technique used in the RTS literature (Bladon et al., 2004; Yang ans Létourneau, 2005; Stratman et al., 2007). Time series analysis typically forecasts a specific variable, given that we know how this variable has changed over time in the past, while in the statistical models such as regression and other predictive techniques, the change of the variables over time component of data is either ignored or is not considered.

A stochastic model (such as gamma process and Markov models) is usually used for estimating probability distributions of potential outcomes. This is done by allowing for random variation based on fluctuations observed in historical data in one or more inputs over time (Berger et al., 2011; Yousefikia et al., 2014; Flier et al., 2009; Sun et al., 2015). As an example, Berger et al. (2011) used a stochastic model for delay propagation and forecast of arrival and departure events of trains as railway traffic always deviates from the planned schedule to a certain extent and the primary initial delays of trains may cause a whole cascade of secondary delays of other trains over the entire network.

Other well-applied prediction models in the context of RTS include regression analysis Zarembski et al. (2016), Murali et al. (2010), Shafiullah et al. (2010) and logit models Tyler Dick et al. (2003), Jinhua (2004), Hajibabai et al. (2012).

Coming to the optimization-based methods, they are mainly categorized into two categories of exact and heuristic methods. Exact methods usually take a form of branching and other forms of exhaustive search, while heuristic methods generally invoke the local search or imitation of a natural process, such as annealing or biological evolution for solving computationally hard optimization problems.

For exact methods, examples in the RTS literature could be found in He et al. (2015), Chen et al. (2010), Su et al. (2016), Sharma et al. (2017). With regard to heuristic methods, Fumeo et al. (2015), Chen et al. (2010), Lee et al. (2016), Van Der Hurk et al. (2015) are the examples of the studies in the context of RTS. In this category, evolutionary algorithms and more specifically genetic algorithm have been among the most popular ones used in the literature (Budai et al., 2009; Bešinović et al., 2013).

Statistical modeling includes any statistic-related techniques from simple descriptive statistics to the application of different statistical distributions on a dataset. These widely exist in the literature (Zhang et al., 2016; Mercier et al., 2012; Zhu et al., 2013; Chattopadhyay and Kumar, 2009; Corbetta et al., 2015; Lin and Saat, 2014).

A key feature of big data is the variety of data sources that are available; which includes not just numerical data but also image or video data or even free text which requires different techniques of image processing (Jamshidi et al., 2017) and text analysis (Zhao et al., 2014; Hughes et al., 2015). Coming to the more advanced models of BDA in RTS, text analysis is one of the most prominent. It is about parsing texts in order to extract machine-readable facts from them. The purpose of text analysis is to create sets of structured data out of unstructured and heterogeneous documents. Natural Language Processing (NLP) is one of the most used techniques related to text analysis. As an example in the context of RTS, Hughes et al. (2015) used the NLP technique to uncover safety information from close calls in the Great Britain Railways. Semantic analysis is also one of the most useful techniques related to text analysis models. Semantic analysis aims to determine the attitude of individuals (the public, specific business customers, transit passengers, etc.) with respect to certain issues (Henry, 2013). Interest in sentiment analysis is expedited by the prominence of social media. The propagation of verbal data including consumer reviews, subjective ratings, and recommendations, together with other types of online verbal expressions, has dramatically increased attention in regard to this aspect of analytics. For railway agencies, sentiment analysis can play a role as a potentially valuable tool to gauge public attitudes toward the agency's services or to monitor for security threats to transit operations or passengers. Despite the popularity of this method, the application of this method in the literature of RTS is still rare.

Process mining combines process models and data mining. Process mining applies specialized data-mining algorithms to event log datasets in order to identify trends, patterns, and details contained in event logs. The application of these models in RTS could be found in Batley et al. (2011) and Kecman and Goverde (2015).

Simulation techniques are widely used in the context of RTS. Examples of these studies could be found in Jiang et al. (2016), De Fabris et al. (2008), Bešinović et al. (2013).

The comprehensive framework of the reviewed articles based on the four layers is all presented in Table 6.

4. Survey results and future direction

Our survey develops a taxonomy framework which aims to provide a complete picture of where BDA has been applied in RTS, at what level of analytics is BDA used in these RTS areas, and what types of BDA models are used in RTS. And what BDA techniques are employed to develop these models. We have gone through a holistic review to find the answer to these questions as presented in the following subsections. The response to each of these question s, not only presents the state of the practice of BDA in RTS, but also depicts the research gaps of the literature and pave the way for the future direction of research in the application of BDA in RTS.

4.1. RTS areas

Maintenance has been the most popular RTS area studied by researchers in the recent years. Considering the types of maintenance in railway, according to the reviewed articles, preventive maintenance receives the highest level of focus. However, if we consider the type of maintenance with respect to the components on which the maintenance is taken over, PM is more popular for track components while CBM is more applied for vehicle components. CM has been mainly studied for tracks but not for vehicle and signaling equipment and few studies have considered the preventive maintenance on other vehicle components rather than wheels.

Considering the component by itself, vehicle maintenance prevails with a particular focus on maintenance of four components: wheel, bearing, truck, and traction. It is worth mentioning that BDA-driven studies for wheel are mainly studied in the condition-based level of maintenance (Enblom and Berg, 2005; Adam et al., 2013; Palo, 2014) while the use of BDA at the corrective level for wheel maintenance is only studied in few studies such as the one by Grassie (2005). Moreover, the application of BDA on truck and bearing maintenance have recently gained more attention, but they are still under-examined in both corrective and preventive levels (Shafiullah et al., 2010). Finally, all types of maintenance for locomotive and traction are seldom addressed (Sammouri et al., 2013; Morgado et al., 2008).

Coming to track maintenance, PM is currently receiving a lot of research interest for both of the main components of track: structure and geometry; The application of BDA theories and tools on this topic is at a relatively mature stage (Hu and Liu, 2016; Yousefikia et al., 2014; Mercier et al., 2012; Zhu et al., 2013; Chattopadhyay and Kumar, 2009; Sadeghi and Askarinejad, 2012; Zarembski et al., 2016). Corrective maintenance has been studied mainly for track structure deterioration models compared to track geometry models. Moreover, according to the discussion on the types of monitoring cars for track maintenance, it is expected that

Table 6
Detailed classification of the articles.

Structure layer		Taxonomy		
RTS Areas	Maintenance	Vehicle (Morgado et al., 2008)	Condition of Wheel	Grassie (2005), Braghin et al. (2006), Sura (2011), Hajibabai et al. (2012), Papaelias et al. (2016), Skarlatos et al. (2004), Yang and Létourneau (2009), Bladon et al. (2004), Enblom and Berg (2005), Adam et al. (2013), Palo (2014), Li and He (2015), Li et al. (2014)
			Truck	Shafiullah et al. (2010), Bladon et al. (2004), Palo (2014) and Li and He (2015)
			Bearing	Papaelias et al. (2016), Enblom and Berg (2005), Loutas et al. (2013), Fumeo et al. (2015) and Vale et al. (2016)
			Tilt and traction	Sammouri et al. (2013) and Morgado et al. (2008)
		Track	Track geometry condition	He et al. (2015), Nunez et al. (2015), Budai et al. (2009), Budai-Balke (2009), Hu and Liu (2016), Zhu et al. (2013), Sadeghi and Askarinejad (2012), Zarembski et al. (2016), Shang and Berenguer (2014), Giben et al. (2015) and Sharma et al. (2017)
			Track structure condition	Tyler Dick et al. (2003), Liu et al. (2006), Schafer and Barkan (2008), Corbetta et al. (2015), Jamshidi et al. (2017), Kumar (2006), Budai et al. (2009), Zhi et al. (2016), Chattopadhyay and Kumar (2009), Sadeghi and Askarinejad (2012), Zarembski et al. (2016), Kaewunruen (2014) and Su et al. (2016)
			Signaling Equipment	Zhao et al. (2014), Zhao et al. (2013), Zhang et al. (2016), Sammouri et al. (2013), Yin and Zhao (2016), Kaewunruen (2014) and Yilboga et al. (2010)
		Type of Maintenance	Corrective Maintenance	Grassie (2005), He et al. (2015), Tyler Dick et al. (2003), Schafer and Barkan (2008), Corbetta et al. (2015), Jamshidi et al. (2017) and Nunez et al. (2015)
			Preventive Maintenance	Braghin et al. (2006), Sura (2011), Hajibabai et al. (2012), Papaelias et al. (2016), Skarlatos et al. (2004), Yang and Létourneau (2009), Shafiullah et al. (2010), Sammouri et al. (2013), Morgado et al. (2008), Budai et al. (2009), Hu and Liu (2016), Yousefikia et al. (2014), Mercier et al. (2012), Zhu et al. (2013), Sadeghi and Askarinejad (2012), Zarembski et al. (2016), Kumar (2006), Budai et al. (2009), Zhi et al. (2016), Chattopadhyay and Kumar (2009), Sadeghi and Askarinejad (2012), Zarembski et al. (2016), Zhao et al. (2013), Zhang et al. (2016) and Sammouri et al. (2013)
			Condition-Based Maintenance	Bladon et al. (2004), Enblom and Berg (2005), Adam et al. (2013), Palo (2014), Li and He (2015), Li et al. (2014), Li and He (2015), Enblom and Berg (2005), Loutas et al. (2013), Fumeo et al. (2015), Vale et al. (2016), Giben et al. (2015), Kaewunruen (2014), Yin and Zhao (2016), Yilboga et al. (2010), Su et al. (2016), Sharma et al. (2017) and Shang and Berenguer (2014)
	Operation	Route Choice		Kusakabe et al. (2010), Van Der Hurk et al. (2015), Sun et al. (2012), Sun et al. (2015) and Jiang et al. (2016)
		Train Positioning and Conflict Detection		Daamen, et al. (2009), Kecman and Goverde (2012), Goverde et al. (2008), Goverde and Meng (2011), De Fabris et al. (2008), Medeossi et al. (2011), Bešinović et al. (2013) and Chen et al. (2010)
		Demand/Market Forecasting, Disruption management		Tsai et al. (2009), Jiang et al. (2014), Sun et al. (2015), Batley et al. (2011) and Wei et al. (2017)
				Zilko et al. (2016)
		Train Delay Estimation		Kecman and Goverde (2015a), Hansen et al. (2010), Flier et al. (2009), Lee et al. (2016), Berger et al. (2011), Wang and Work (2015), Yaghini et al. (2013), Markovic et al. (2015), Oneto et al. (2016), Murali et al. (2010) and Goverde and Meng (2011)
	Safety	Factor Analysis and Frequency Analysis		Mok and Savage (2005), Evans (2010), Silla and Kallberg (2012), Kyriakidis et al. (2012), Lin and Saat (2014), Liu (2017), Liu et al. (2017), Liu (2016a) and Liu (2016b)
		Severity Analysis		Liu et al. (2013), Ghomi et al. (2016)
		Risk Analysis		Liu et al. (2011), Liu et al. (2017), Liu (2016a), Mirabadi and Sharifian (2010), Liu et al. (2012), Shao et al. (2016) and Baysari et al. (2008)

(*continued on next page*)

Table 6 (*continued*)

Structure layer		Taxonomy	
Analytics Level (Delen and Demirkan, 2013)	Descriptive	Enblom and Berg (2005), Sammouri et al. (2013), Morgado et al. (2008), Zhu et al. (2013), Zarembski et al. (2016), Grassie (2005), Bearfield and Marsh (2010), Mirabadi and Sharifian (2010), Evans (2010), Liu et al. (2012), Wu et al. (2015), Ghomi et al. (2016), Shao et al. (2016), Baysari et al. (2008), Evans (2011), Kyriakidis et al. (2012), Evans (2011), Batley et al. (2011), Daamen et al. (2009), Medeossi et al. (2011), Kecman and Goverde (2015a), Sun et al. (2012), Liu et al. (2011), Liu et al. (2017), Liu (2016a), Liu et al. (2013), Liu (2016b), Liu et al. (2014), Bagheri et al. (2014), Liu (2016c), Castillo et al. (2016), Goverde and Meng (2011) and Goverde et al. (2008)	
	Predictive	Tyler Dick et al. (2003), Yang and Létourneau (2005), Kumar (2006), Braghin et al. (2006), Liu et al. (2006), Yang and Létourneau (2009), Budai et al. (2009), Yilboga et al. (2010), Shafiullah et al. (2010), Sura (2011), Hajibabai et al. (2012), Zhao et al. (2013), Bevan et al. (2013), Fumeo et al. (2015), Giben et al. (2015), Yin and Zhao (2016), Yousefikia et al. (2014), Mercier et al. (2012), He et al. (2015), Zhi et al. (2016), Chattopadhyay and Kumar (2009), Sadeghi and Askarinejad (2012), Skarlatos et al. (2004), Li et al. (2014), Bladon et al. (2004), Jamshidi et al. (2017), Hansen et al. (2010), Murali et al. (2010), Berger et al. (2011), Wang and Work (2015), Kecman and Goverde (2015a), Markovic et al. (2015), Oneto et al. (2017b), Tsai et al. (2009), De Fabris et al. (2008), Flier et al. (2009), Lee et al. (2016), Oneto et al. (2016), Oneto et al. (2017a), Jiang et al. (2014), Sun et al. (2015) and Zilko et al. (2016)	
	Prescriptive	Bladon et al. (2004), Oneto et al. (2017a), Starke et al. (2006), Stratman et al. (2007), Schlake et al. (2010), Sammouri et al. (2013), Loutas et al. (2013), Zhao et al. (2014), Palo (2014), Li and He (2015), Hu and Liu (2016), Papaelias et al. (2016), Zhang et al. (2016), Vale et al. (2016), Schafer and Barkan (2008), Corbetta et al. (2015), Chen et al. (2010), Kusakabe et al. (2010), Liu et al. (2011), Bearfield et al. (2013), Lin and Saat (2014), Jiang et al. (2016), Hughes et al. (2015), Wei et al. (2017), Bešinović et al. (2013), Van Der Hurk et al. (2015), Sun et al. (2015), Sharma et al. (2017), Su et al. (2016), Nunez et al. (2015) and Shang and Berenguer (2014)	
BDA Models and Techniques (Erl, Khattak and Buhler, 2016)	Association	Mirabadi and Sharifian (2010), Sammouri et al. (2013), Ghomi et al. (2016), Vale et al. (2016), Zarembski et al. (2016), Li et al. (2014), Mirabadi and Sharifian (2010), Liu et al. (2011), Liu et al. (2012) and Ghomi et al. (2016)	
	Clustering	Hughes et al. (2015), Shao et al. (2016) and Su et al. (2016)	
	Simulation	Enblom and Berg (2005), Braghin et al. (2006), Sura (2011), Bevan et al. (2013), Palo (2014), Vale et al. (2016), Zhi et al. (2016), Corbetta et al. (2015), Jiang et al. (2016), De Fabris et al. (2008), Bešinović et al. (2013), Sun et al. (2015), Sharma et al. (2017) and Shang and Berenguer (2014)	
	Image Processing	Schlake et al. (2010), Giben et al. (2015), Jamshidi et al. (2017) and Zhu et al. (2013)	
	Process Mining	Batley et al. (2011), Daamen et al. (2009) and Kecman and Goverde (2012)	
	Statistical Analysis	Zhang et al. (2016), Mercier et al. (2012), Chattopadhyay and Kumar (2009), Corbetta et al. (2015), Zarembski et al. (2016), Skarlatos et al. (2004), Grassie (2005), Goverde and Meng (2011), Nunez et al. (2015), Zarembski et al. (2016), Murali et al. (2010), Shafiullah et al. (2010), Tyler Dick et al. (2003), Jinhua (2004), Hajibabai et al. (2012), Li and He (2015), Schafer and Barkan (2008), Zarembski et al. (2016), Li et al. (2014), Wang and Work (2015), Oneto et al. (2017a), Kecman and Goverde (2015a), Oneto et al. (2016) and Sun et al. (2012)	
	Semantic Analysis	Zhao et al. (2014), Giben et al. (2015) and Liu (2016a)	
	Optimization	Exact Methods	He et al. (2015), Chen et al. (2010), Su et al. (2016) and Sharma et al. (2017)
		Heuristic Methods	Fumeo et al. (2015), Chen et al. (2010), Lee et al. (2016), Van Der Hurk et al. (2015), Budai et al. (2009) and Bešinović et al. (2013)
	Prediction	Classification	Yang and Létourneau (2009), Yin and Zhao (2016), Ghomi et al. (2016), Yang and Létourneau (2005), Zhao et al. (2014), Giben et al. (2015), Bearfield and Marsh (2010), Bearfield et al. (2013), Wu et al. (2015), Kecman and Goverde (2015a) and Lee et al. (2016)
		Pattern Recognition Models, Artificial Neural Network (ANN),	Yu et al. (2007), Yilboga et al. (2010) and Chen and Gao (2012)
		Support Vector Machin (SVM),	Hu and Liu (2016)
		Support Vector Regression (SVR),	Loutas et al. (2013), Fumeo et al. (2015) and Markovic et al. (2015)
		Bayesian Network (BN),	Bearfield et al. (2013), Sun et al. (2015) and Zilko et al. (2016)
		Time Series Models	Bladon et al. (2004), Yang and Létourneau (2005) and Stratman et al. (2007)

drone-based inspection including the fast image processing and analysis would be emerging in the near future. CM of the geometry components of the track as well as the CBM for both geometry and structure components of the track have a very limited number of studies in the literature He et al. (2015), Giben et al. (2015) and Kaewunruen (2014).

Operation is the second major area of RTS which can profit from BDA. Note that we define the scope of operations in our study as the activities related to train traffic and transportation services. With regards to the operations management, it is well documented in the literature that traffic control decisions on train positioning and conflict detection, demand/market forecasting, disruption management and train delay estimation, have taken advantage of BDA Kecman and Goverde (2012), Jiang et al. (2014), Zilko et al. (2016) and Kecman and Goverde (2015a). However, data about disruptions are very poorly reported and often do not include much more than the failing element in free-form text fields in rail traffic control information systems, while information on the underlying failing component and required repair are not usually available. Further, very few studies of BDA in this domain can be found in the literature (Zilko et al., 2016). One of the main challenges in this field of study is privacy and confidentiality of data. Smart card data can be used to derive the travel behavior of individual persons. Train delay data for specific trains can be traced back to train operators, individual drivers, and dispatchers; And likewise, disruption repair data can be traced to contractors. Although world wide web is a good source of railway operations data, publications using this data are few in the literature (Wei et al., 2017), but this is a promising area as more and more train operating companies increasingly share dynamic data on their websites.

Although BDA adoption in safety as the third main area of RTS is relatively less studied, papers in this area make a significant contribution to descriptive analytics. On the other hand, research on BDA-enabled predictive and prescriptive studies is relatively limited. BDA studies on safety are usually analyses in terms of influencing factors, accident cause, frequency, accident severity as well as risk of accident (Lin and Saat, 2014; Ghomi et al., 2016; Baysari et al., 2008). The studies on these perspectives almost evenly spread over the literature. However, due to the nature of safety problems in the context of RTS, probabilistic models for risk analysis of accidents are focused mainly in this field of study, specifically those rare accidents with very severe consequences (Liu et al., 2011, 2017; Liu, 2016a). There are numerous big data sources in the context of RTS including monitoring the rail infrastructure and train operations. However, the main challenge is the procedure on how to leverage these big data sources for safety analysis which requires a deep understanding of their relationships with safety (Bearfield et al., 2013). Finally, the review shows that at present, the linkage between small accident data and other big data sources is not well understood. Understanding this would not only demand advanced analytical approaches but also a practical understanding of various engineering, operational and management aspects of RTS.

4.2. Level of analytics

The aim of this question is to investigate the level of BDA required in the RTS application, as well as indicating the types of problem being solved. According to the results of the trend analysis for the level of analytics, the application of data analysis in RTS is mainly at descriptive level which is closely followed by predictive analytics, while prescriptive analytics is receiving less consideration. To be more specific, safety is the major contributor of the descriptive analysis, mainly due to the small datasets available in this area which limits the analysis at descriptive level. On the other hand, predictive analytics is the primary actor in operations thanks to the importance and widespread use of train delay prediction and train conflict detection in this field of study (Kecman and Goverde, 2015a). For the maintenance area, prescriptive analysis plays a role although it is still dominated by predictive analysis. The main reason for this is the importance of predictive and prescriptive analysis of maintenance-related data in reducing the cost imposed to the systems before a failure occurs. Finally, the review shows that prescriptive analytics in the safety and operations areas are rarely discussed.

4.3. BDA models and techniques

As we discussed in Section 3.3, researchers use several approaches in the literature for analysis of data in different problems of RTS. Association and clustering are of the most popular approaches in the realm of descriptive analytics. These two models and their corresponding techniques (more particularly association rules and K-means) are specifically applied for maintenance and safety areas of RTS. Few studies have considered the descriptive level in the operations area.

At the predictive level of analytics, classification, pattern recognition models, time series analysis, and stochastic models are well applied in the studied papers among which pattern recognition models are the most popular models. ANN and SVM are two of the most popular techniques in this domain applied in the RTS studies. These models are mostly considered in the maintenance and operations areas. Studies on the predictive level of safety in RTS are almost scarce in the literature.

At the prescriptive level, optimization-based models prevail in two main categories of exact and heuristic models among which GA has been the most used technique in the RTS literature.

Through the progression of computer science technologies for data collection and processing, newer advanced methods are now available to be applied in the context of RTS. These are discussed further in the following subsection.

4.3.1. BDA advanced methods: The state of the practice in RTS

On the advanced methods, interests in sentiment analysis and language processing are being expedited by the prominence of social media platforms such as Twitter. The task usually involves detecting whether a piece of text expresses a POSITIVE, a NEGATIVE, or a NEUTRAL sentiment. The application of sentiment analysis for the transportation industry has been just considered in the very recent years but not very frequently in the literature. Most of these transportation-related studies have focused particularly on “traffic sentiment analysis” to meet the needs of safety and information exchange in intelligent transportation systems (Cao et al.,

2014; Ali et al., 2017; Zhang et al., 2017, 2018). However, considering railway as a specific mode of transportation, sentiment analysis studies are very rare which implies a great opportunity for researchers to consider this field of study which is of great interest for future studies of BDA in the context of RTS. It is estimated that for railway agencies, sentiment analysis can play a role as a potentially valuable tool to gauge public attitudes toward the agency's services or to monitor for security threats to transit operations or passengers.

On the other side, data collection in the RTS context by itself is usually accompanied by several challenges. First of all, in most cases, the data collected suffers from heterogeneity, inconsistency, and incompleteness. This is mainly due to the reason that RTS data is usually collected in different formats from different sources in different spatial and geographical locations. As an example, the data collected during track inspection includes both rail and geometry defect data as well as tonnage data. On the other hand, collecting track data in real time especially for maintenance purposes may have a time limit, which requires real-time data analytic techniques (Summit, 2014). Analyzing big data in railroads is not only about large databases but it also includes merging of different databases to extract information for further analysis. Another important issue of BDA for RTS is associated with the issue of privacy and data ownership. For transit agencies, it is possible to extract great volumes of big data from fare transaction data, passenger counts, etc. All of this provides the opportunity for abuse which is a major concern within the RTS professionals that make the railway companies more conservative in data sharing. So far, most of the techniques and models that have been used in RTS are not really dealing with very large datasets. The huge amount of generated data in RTS necessitates the application of newer technologies/tools with capabilities in handling this data. In this regard, the tools that are currently developed in computer science can yield useful applications for different domains of railways industry. However, it requires reliable and dedicated computer systems as well as novel data analytic techniques, linguistic tools, and sensible interface-techniques. All of these require dedicated collaboration projects between railway engineers, scientists, information technologists and software developers.

One of the newest data resources commonly used for many BDA applications is Apache Hadoop (which is based on cloud computing), a free software which distributes data and algorithms over several computers and collects the results after they are processed White (2015). A computational paradigm named MapReduce is implemented by Hadoop by which the application task is divided into many small fragments of work jobs (Zhang et al., 2009). MapReduce runs on a distributed file system (HDFS) that stores data on the computer nodes providing very high aggregate bandwidth.

Tashi Cloud Middleware is also a relatively new cluster management system for cloud computing on big data. Tashi is designed to support cloud computing applications that operate on Big Data. It is a virtualization-based cluster management system that provides facilities for managing virtual machines. Users of Tashi are able to create collections of virtual machines that run on the cluster's physical resources (Zhang et al., 2009). China Ministry of Railway implemented a novel cloud computing-based freight system application for a freight query & tracking service based on the HDFS and MapReduce mechanism. They used Tashi and Hadoop to implement the application system. It includes a freight ticket sub-system, a confirmation report sub-system, and a train dispatching sub-system. When the freight is grouping to a specific wagon for a next railway station and the freight wagon arrives at the new railway station, the confirmation report information is made for affirming the arrival of freight. They utilized the Tashi cloud middleware to manage the virtual machine resource, including CPU, memory, and storage space of every physical machine within cluster management. The Hadoop distributed files system (HDFS) provides the data management mechanism. Based on the HDFS, a MapReduce distributed programming model provides a parallel search mechanism for processing the large-scale data.

The GB Railways is also one of the pioneers in investigating the extent to which big data technologies can support railway safety issues and risk analysis (Van Gulijk et al., 2015). They combined three major components, a Hadoop computer cluster, industry servers containing databases, interface devices and the Internet for big data risk analysis. The GB railway system and many other railway systems depend on the collaboration of many organizations in the railway industry including train operating companies, infrastructure managers, rolling stock companies, maintenance and construction companies, enforcement bodies, regulatory bodies and many more (Van Gulijk et al., 2015).

5. Conclusion

Based on the content analysis methodology of Mayring (2003), this survey examined 115 articles to provide a comprehensive picture on where and how BDA has been applied in the context of RTS. In particular, we develop a classification framework in four layers: the RTS areas where BDA has been applied, the level of analytics at which BDA has been studied, BDA models, and BDA techniques applied in the context of RTS. By addressing these four aspects, a number of research gaps, future directions, and challenges for BDA applications is highlighted to catalyze the research development of the topic in the future.

References

Adam, B., Molyneux-Berry, P., Eickoff, Bridget, Burstow, M., 2013. Development and validation of a wheel wear and rolling contact fatigue damage mode. Wear 307 (1–2), 100–111.

Al-Jarrah, O.Y., Yoo, P.D., Muhaidatc, S., Karagiannidis, G.K., Tahaa, K., 2015. Efficient machine learning for big data: a review. Big Data Res. 2 (3), 87–93.

Ali, F., Kwak, D., Khan, P., Islam, S.M.R., Kim, K.H., Kwak, K.S., 2017. Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling. Transport. Res. Part C: Emerg. Technol. 77, 33–48.

Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A.S., Buyya, R., 2015. Big data computing and clouds: trends and future directions. J. Parallel Distrib. Comput. 79–80 (May), 3–15.

Bagheri, M., Verma, M., Verter, V., 2014. Transport mode selection for toxic gases: rail or road? Risk Anal. 34 (1), 168–186.

Batley, R., Dargay, J., Wardman, M., 2011. The impact of lateness and reliability on passenger rail demand. Transport. Res. Part E: Logist. Transport. Rev. 47 (1), 61–72.

Baysari, M.T., McIntosh, A.S., Wilson, J.R., 2008. Understanding the human factors contribution to railway accidents and incidents in Australia. Accid. Anal. Prev. 40 (5), 1750–1757.
Bearfield, G., Marsh, W., 2010. Causal modelling of lower consequence rail safety incidents. Reliability, Risk and Safety: Back to the Future 2010(Esrel), 2261–2266.
Bearfield, G., Holloway, A., Marsh, W., 2013. Change and safety: decision-making from data. Proc. Inst. Mech. Eng. Part F: J. Rail Rapid Transit 227 (6), 704–714.
Berger, A., Gebhardt, A., Müller-Hannemann, M., Ostrowski, M., 2011. Stochastic delay prediction in large train networks. In: 11th Workshop on Algorithmic Approaches for Transportation Modelling Optimization, and Systems, vol. 20, pp. 100–111.
Bešinović, N., Quaglietta, E., Goverde, R.M.P., 2013. A simulation-based optimization approach for the calibration of dynamic train speed profiles. J. Rail Transp. Plann. Manage. 3 (4), 126–136.
Bevan, A., Molyneux-Berry, P., Eickhoff, B., Burstow, M., 2013. Development and validation of a wheel wear and rolling contact fatigue damage model. Wear 307 (1–2), 100–111.
Bladon, K., Rennison, D., Izbinsky, G., Tracy, R., Bladon, T., 2004. Predictive condition monitoring of railway rolling stock. In: CORE 2004: New Horizons for Rail. Darwin, N.T.: Railway Technical Society of Australasia, pp. 1–22.
Braghin, F., Lewis, R., Dwyer-Joyce, R.S., Bruni, S., 2006. A mathematical model to predict railway wheel profile evolution due to wear. Wear 261 (11–12), 1253–1264.
Brandenburg, M., Govindan, K., Sarkis, J., Seuring, S., 2014. Quantitative models for sustainable supply chain management: developments and directions. Eur. J. Oper. Res. 233 (2), 299–312.
Brewerton, P.M., Millward, L.J., 2001. Organizational Research Methods: A Guide for Students and Researchers. Sage.
Budai-Balke, G., 2009. Operations Research Models for Scheduling Railway Infrastructure Maintenance. PhD Thesis, Erasmus University Rotterdam.
Budai, G., Dekker, R., Kaymak, U., 2009. Genetic and Memetic Algorithms for Scheduling Railway Maintenance Activities. Erasmus University Rotterdam, No. EI 200, pp. 1–23.
Budai, G., Dekker, R., Nicolai, R.P., 2006. A Review of Planning Models for Maintenance & Production. Erasmus University Rotterdam.
Cao, J., Zeng, K., Wang, H., Cheng, J., Qiao, F., Wen, D., Gao, Y., 2014. Web-based traffic sentiment analysis: methods and applications. IEEE Trans. Intell. Transport. Syst. 15 (2), 844–853.
Castillo, E., Grande, Z., Calvino, A., 2016. Bayesian networks-based probabilistic safety analysis for railway lines. Comput.-Aid. Civil Infrastruct. Eng. 31 (9), 681–700.
Chattopadhyay, G., Kumar, S., 2009. Parameter estimation for rail degradation model. Int. J. Perform. Eng. 5 (2), 119–130.
Chen, D., Fu, Y.S., Cai, B., Yuan, Y.X., 2010. Modeling and algorithms of GPS data reduction for the Qinghai-Tibet railway. IEEE Trans. Intell. Transp. Syst. 11 (3), 753–758.
Chen, D., Gao, C., 2012. Soft computing methods applied to train station parking in urban rail transit. Appl. Soft Comput. J. 12 (2), 759–767.
Corbetta, M., Sbarufatti, C., Manes, A., Giglio, M., 2015. Real-time prognosis of crack growth evolution using sequential Monte Carlo methods and statistical model parameters. IEEE Trans. Reliab. 64 (2), 736–753.
Daamen, W., Goverde, R.M.P., Hansen, I.A., 2009. Non-discriminatory automatic registration of knock-on train delays. Networks Spatial Econ. 9 (1), 47–61.
De Fabris, S., Longo, G., Medeossi, G., 2008. Automated analysis of train event recorder data to improve micro-simulation models. WIT Trans. Built Environ. 103, 575–583.
Delen, D., Demirkan, H., 2013. Data, information and analytics as services. Decis. Support Syst. 55 (1), 359–363.
Duan, L., Xiong, Y., 2015. Big data analytics and business analytics. J. Manage. Anal. 2 (1), 1–21.
Enblom, R., Berg, M., 2005. Simulation of railway wheel profile development due to wear influence of disc braking and contact environment. Wear 258 (7–8), 1055–1063.
Erl, T., Khattak, W., Buhler, P., 2016. Big Data Fundamentals: Concepts, Drivers & Techniques. Prentice Hall Press.
Evans, A.W., 2010. Rail safety and rail privatisation in Japan. Accid. Anal. Prev. 42 (4), 1296–1301.
Evans, A.W., 2011. Fatal train accidents on Europe's railways: 1980–2009. Accid. Anal. Prev. 43 (1), 391–401.
Evans, A.W., Verlander, N.Q., 1996. Estimating the consequences of accidents: the case of automatic train protection in Britain. Accid. Anal. Prev. 28 (2), 181–191.
Faulkner, A., 2002. Safer data: the use of data in the context of a railway control system. In: Components of System Safety. Springer, pp. 217–230.
Figueres-Esteban, M., Hughes, P., Van Gulijk, C., 2015. The role of data visualization in railway Big Data Risk Analysis. In: 25th European Safety and Reliability Conference, ESREL 2015, pp. 2877–2882.
Figueroa-García, J.C., Kalenatic, D., López-Bello, C.A., 2015. Intelligent Techniques in Engineering Management. Intelligent Systems Reference Library, vol. 87.
Flier, H., Gelashvili, R., Graffagnino, T., Nunkesser, M., 2009. Mining railway delay dependencies in large-scale real-world delay data. Robust Online Large-Scale Optim. 5868, 354–368.
Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., Gnanzou, D., 2015. How "big data" can make big impact: findings from a systematic review and a longitudinal case study. Int. J. Prod. Econ. 165, 234–246.
Fumeo, E., Oneto, L., Anguita, D., 2015. Condition based maintenance in railway transportation systems based on big data streaming analysis. Proc. Comput. Sci. 53 (1), 437–446.
Gandomi, A., Haider, M., 2015. Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manage. 35 (2), 137–144.
Ghomi, H., Bagheri, M., Fu, L., Miranda-Moreno, L.F., 2016. Analysing injury severity factors at highway railway grade crossing accidents involving vulnerable road users: a comparative study. Traffic Inj. Prev. 9588 (May).
Giben, X., Patel, V.M., Chellappa, R., 2015. Material classification and semantic segmentation of railway track images with deep convolutional neural networks. In: Proceedings - International Conference on Image Processing, ICIP, 2015–Decem (June), pp. 621–625.
Gläser, J., Laudel, G., 2010. Experteninterviews und qualitative Inhaltsanalyse. Springer-Verlag.
Global Railway Review, 2017. Big data in railway operations and maintenance. Global Railway Review magazine, Aug 10, 2017 (https://www.globalrailwayreview.com/article/61515/big-data-railway-operations-maintenance-2/).
Goverde, R., Hansen, I.A., 2000. TNV-prepare: analysis of Dutch railway opeartions based on train detection data. Comput. Railways VII 50, 10.
Goverde, R.M.P., Daamen, W., Hansen, I.A., 2008. Automatic identification of route conflict occurrences and their consequences. WIT Trans. Built Environ. 103, 473–482.
Goverde, R.M.P., Meng, L., 2011. Advanced monitoring and management information of railway operations. J. Rail Transp. Plann. Manage. 1 (2), 69–79.
Govindan, K., Soleimani, H., Kannan, D., 2015. Reverse logistics and closed-loop supply chain: a comprehensive review to explore the future. Eur. J. Oper. Res. 240 (3), 603–626.
Grassie, S.L., 2005. Rolling contact fatigue on the British railway system: treatment. Wear 258 (7–8), 1310–1318.
Griffin, D.W.P., Mirza, O., Kwok, K., Kaewunruen, S., 2014. Composite slabs for railway construction and maintenance: a mechanistic review. IES J. Part A: Civil Struct. Eng. 7 (4), 243–262.
Hajibabai, L., Saat, M.R., Ouyang, Y., Barkan, C.P.L., Yang, Z., Bowling, K., Somani, K., Lauro, D., Li, X., 2012. Wayside defect detector data mining to predict potential WILD train stops. In: Proceedings of American Railway Engineering and Maintenance-of-Way Association Annual Meeting.
Hansen, I.A., Goverde, R.M.P., Van Der Meer, D.J., 2010. Online train delay recognition and running time prediction. In: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, pp. 1783–1788.
He, Q., Li, H., Bhattacharjya, D., Parikh, D.P., Hampapur, A., 2015. Track geometry defect rectification based on track deterioration modelling and derailment risk assessment. J. Oper. Res. Soc. 66 (3), 392–404.
Henry, J., 2013. Analytics and big data-rail public transportation is a leader. In: Rail Conference of the American Public Transportation Association.
Hodge, V.J., O'Keefe, S., Weeks, M., Moulds, A., 2015. Wireless sensor networks for condition monitoring in the railway industry: a survey. IEEE Trans. Intell. Transp. Syst. 16 (3), 1088–1106.
Hu, C., Liu, X., 2016. Modeling track geometry degradation using support vector machine technique. In: 2016 Joint Rail Conference. American Society of Mechanical Engineers, pp. V001T01A011–V001T01A011.

Hughes, P., Van Gulijk, C., Figueres-Esteban, M., 2015. Learning from text-based close call data. In: Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015, 7353(June), 8.
Jamshidi, A., Faghih-Roohi, S., Hajizadeh, S., Núñez, A., Babuska, R., Dollevoet, R., De Schutter, B., 2017. A big data analysis approach for rail failure risk assessment. Risk Anal. 37 (8).
Jardine, A.K.S., Lin, D., Banjevic, D., 2006. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech. Syst. Sig. Process. 20 (7), 1483–1510.
Jiang, X., Zhang, L., Chen, M.X., 2014. Short-term forecasting of high-speed rail demand: a hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in China. Transport. Res. Part C: Emerg. Technol. 44 (August 2015), 110–127.
Jiang, Z., Hsu, C.H., Zhang, D., Zou, X., 2016. Evaluating rail transit timetable using big passengers' data. J. Comput. Syst. Sci. 82 (1), 144–155.
Jinhua, Z., 2004. The Planning and Analysis Implications of Automatic Data Collection Systems: Rail Transit OD Matrix Inference and Path Choice Modelling Examples. Thesis of Master, M.S. Thesis, MIT, USA.
Emani, C.Kacfah, Cullot, N., Nicolle, C., 2015. Understandable Big Data: a survey. Comput. Sci. Rev. 17, 70–81.
Kaewunruen, S., 2014. Monitoring structural deterioration of railway turnout systems via dynamic wheel/rail interaction. Case Stud. Nondestr.Test. Eval. 1, 19–24.
Kecman, P., Goverde, R.M.P., 2012. Process mining of train describer event data and automatic conflict identification. WIT Trans. Built Environ. 127, 227–238.
Kecman, P., Goverde, R.M.P., 2015a. Online data-driven adaptive prediction of train event times. IEEE Trans. Intell. Transp. Syst. 16 (1), 465–474.
Kecman, P., Goverde, R.M.P., 2015b. Predictive modelling of running and dwell times in railway traffic. Public Transport 7 (3), 295–319.
Klewitz, J., Hansen, E.G., 2014. Sustainability-oriented innovation of SMEs: a systematic review. J. Cleaner Prod. 65, 57–75.
Krippendorff, K., 2004. Content Analysis: An Introduction to its Methodology. Sage.
Kumar, S., 2006. A Study of the Rail Degradation Process to Predict Rail Breaks, 93.
Kusakabe, T., Iryo, T., Asakura, Y., 2010. Estimation method for railway passengers' train choice behavior with smart card transaction data. Transportation 37 (5), 731–749.
Kyriakidis, M., Hirsch, R., Majumdar, A., 2012. Metro railway safety: an analysis of accident precursors. Saf. Sci. 50 (7), 1535–1548.
Lee, W.H., Yen, L.H., Chou, C.M., 2016. A delay root cause discovery and timetable adjustment model for enhancing the punctuality of railway services. Transport. Res. Part C: Emerg. Technol. 73, 49–64.
Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D., Hampapur, A., 2014. Improving rail network velocity: a machine learning approach to predictive maintenance. Transport. Res. Part C: Emerg. Technol. 45, 17–26.
Li, Z., He, Q., 2015. Prediction of railcar remaining useful life by multiple data source fusion. IEEE Trans. Intell. Transp. Syst. 16 (4), 2226–2235.
Lin, C., Saat, M.R., 2014. Semi-quantitative risk assessment of adjacent track accidents on shared-use rail corridors. In: Proceedings of the 2014 Joint Rail Conference, vol. 3773, pp. 1–10.
Liu, X., 2016a. Analysis of collision risk for freight trains in the United States. Transport. Res. Rec.: J. Transport. Res. Board 2546, 121–128.
Liu, X., 2016b. Statistical causal analysis of freight-train derailments in the United States. J. Transport. Eng., Part A: Syst. 143 (2), 4016007.
Liu, X., 2016c. Statistical temporal analysis of freight-train derailment rates in the United States: 2000 to 2012. Federal Rail Assoc. (848), 1–18.
Liu, X., 2015. Statistical temporal analysis of freight-train derailment rates in the United States : 2000 to 2012. Transport. Res. Rec.: J. Transport. Res. Board 2476, 119–125.
Liu, X., 2017. Statistical causal analysis of freight-train derailments in the United States. J. Transport. Eng., Part A: Syst. 143 (2), 1–8.
Liu, X., Hong, Y.L., Liu, C., 2017a. Analysis of multiple tank car releases in train accidents. Accid. Anal. Prev. 107, 164–172.
Liu, X., Barkan, C., Saat, M., 2011. Analysis of derailments by accident cause: evaluating railroad track upgrades to reduce transportation risk. Transport. Res. Rec.: J. Transport. Res. Board 2261, 178–185.
Liu, X., Rapik Saat, M., Barkan, C.P.L., 2017b. Freight-train derailment rates for railroad safety and risk analysis. Accid. Anal. Prev. 98, 1–9.
Liu, X., Saat, M., Barkan, C., 2012. Analysis of causes of major train derailment and their effect on accident rates. Transport. Res. Rec.: J. Transport. Res. Board 2289, 154–163.
Liu, X., Saat, M.R., Barkan, C.P.L., 2014. Probability analysis of multiple-tank-car release incidents in railway hazardous materials transportation. J. Hazard. Mater. 276, 442–451.
Liu, X., Saat, M.R., Qin, X., Barkan, C.P.L., 2013. Analysis of U.S. freight-train derailment severity using zero-truncated negative binomial regression and quantile regression. Accid. Anal. Prev. 59, 87–93.
Liu, Y., Stratman, B., Mahadevan, S., 2006. Fatigue crack initiation life prediction of railroad wheels. Int. J. Fatigue 28 (7), 747–756.
Loutas, T.H., Roulias, D., Georgoulas, G., 2013. Remaining useful life estimation in rolling bearings utilizing data-driven probabilistic E-support vectors regression. IEEE Trans. Reliab. 62 (4), 821–832.
Markovic, N., Milinkovic, S., Tikhonov, K.S., Schonfeld, P., 2015. Analyzing passenger train arrival delays with support vector regression. Transport. Res. Part C: Emerg. Technol. 56, 251–262.
Mayring, P., 1990. Einführung in die qualitative Sozialforschung. Eine Anleitung zu qualitativem Denken. Psychologie-Verl. Union, München.
Mayring, P., 2003. Qualitative inhaltsanalyse [Qualitative content analysis]. Qualitative Forschung Ein Handbuch (Qualitative Research: A Handbook), pp. 468–475.
Medeossi, G., Longo, G., de Fabris, S., 2011. A method for using stochastic blocking times to improve timetable planning. J. Rail Transp. Plann. Manage. 1 (1), 1–13.
Mercier, S., Meier-Hirmer, C., Roussignol, M., 2012. Bivariate Gamma wear processes for track geometry modelling, with application to intervention scheduling. Struct. Infrastruct. Eng. 8 (4), 357–366.
Mirabadi, A., Sharifian, S., 2010. Application of association rules in Iranian Railways (RAI) accident data analysis. Saf. Sci. 48 (10), 1427–1435.
Mok, S.C., Savage, I., 2005. Why has safety improved at rail-highway grade crossings? Risk Anal. 25 (4), 867–881.
Morant, A., Galar, D., Tamarit, J., 2012. Cloud computing for maintenance of railway signalling systems. In: Proceedings of the Ninth International Conference on Condition Monitoring and Machinery Failure Prevention Technologies, pp. 551–559.
Morgado, T.L.M., Branco, C.M., Infante, V., 2008. A failure study of housing of the gearboxes of series 2600 locomotives of the Portuguese Railway Company. Eng. Fail. Anal. 15 (1–2), 154–164.
Murali, P., Dessouky, M., Ordóñez, F., Palmer, K., 2010. A delay estimation technique for single and double-track railroads. Transport. Res. Part E: Logist. Transport. Rev. 46 (4), 483–495.
Nguyen, T., Zhou, L., Spiegler, V., Ieromonachou, P., Lin, Y., 2017. Big data analytics in supply chain management: a state-of-the-art literature review. Comput. Oper. Res.
Nunez, A., Hendriks, J., Li, Z., De Schutter, B., Dollevoet, R., 2015. Facilitating maintenance decisions on the Dutch railways using big data: the ABA case study. In: Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014, pp. 48–53.
Nunez, S.G., Attoh-Okine, N., 2015. Metaheuristics in big data: an approach to railway engineering. In: Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014, pp. 42–47.
One, P., Pérez, M., 2003. Rail Defect Identification Handbook, (November), 2004.
Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Anguita, D., 2016. Advanced analytics for train delay prediction systems by including exogenous weather data. In: Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, pp. 458–467.
Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Anguita, D., 2017a. Dynamic delay predictions for large-scale railway networks: deep and shallow extreme learning machines tuned via thresholdout. IEEE Trans. Syst. Man Cybern.: Syst. 1–14.
Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Anguita, D., 2017b. Train delay prediction systems: a big data analytics perspective. Big Data Res. 1, 1–11.
Pall, E., Mathe, K., Tamas, L., Busoniu, L., 2014. Railway track following with the AR.Drone using vanishing point detection. In: Proceedings of 2014 IEEE International Conference on Automation, Quality and Testing, Robotics, AQTR 2014.
Palo, M., 2014. DOC TOR A L T H E S I S Condition-Based Maintenance for Effective and Efficient Rolling Stock Capacity Assurance Mikael Palo. Luleå tekniska

universitet.
Papaelias, M., Amini, A., Huang, Z., Vallely, P., Dias, D.C., Kerkyras, S., 2016. Online condition monitoring of rolling stock wheels and axle bearings. Proc. Inst. Mech. Eng., Part F: J. Rail Rapid Transit 230 (3), 709–723.
Podofillini, L., Zio, E., Vatn, J., 2006. Risk-informed optimisation of railway tracks inspection and maintenance procedures. Reliab. Eng. Syst. Saf. 91 (1), 20–35.
Progressive Railroading, 2017. The rail industry is learning to analyze data to answer specific MOW questions. Progressive Railroading magazine, Feb 2017 (http://www.progressiverailroading.com/csx_transportation/article/IoT-The-rail-industry-is-learning-to-analyze-data-to-answer-specific-MOW-questions–50785).
Rehman, M.H.U., Chang, V., Batool, A., Wah, T.Y., 2016. Big data reduction framework for value creation in sustainable enterprises. Int. J. Inf. Manage. 36 (6), 917–928.
Sadeghi, J., Askarinejad, H., 2012. Application of neural networks in evaluation of railway track quality condition. J. Mech. Sci. Technol. 26 (1), 113–122.
Sammouri, W., Côme, E., Oukhellou, L., Aknin, P., 2013a. Mining floating train data sequences for temporal association rules within a predictive maintenance framework. In: Industrial Conference on Data Mining. Springer, pp. 112–126.
Sammouri, W., Come, E., Oukhellou, L., Aknin, P., Fonlladosa, C.-E., 2013. Floating train data systems for preventive maintenance: a data mining approach. In: Proceedings of 2013 International Conference on Industrial Engineering and Systems Management (IESM), (October), pp. 1–7.
Sammouri, W., Come, E., Oukhellou, L., Aknin, P., Fonlladosa, C.-E., 2013c. Floating train data systems for preventive maintenance: a data mining approach. In: Proceedings of 2013 International Conference on Industrial Engineering and Systems Management (IESM). IEEE, pp. 1–7.
Schafer, D.H., Barkan, C.P.L., 2008. A prediction model for broken rails and an analysis of thier economic impact. In: Proceedings of the American Railway Engineering and Maintenance-of-Way Association Annual Conference, 252(847).
Schlake, B.W., Todorovic, S., Edwards, J.R., Hart, J.M., Ahuja, N., Barkan, C.P.L., 2010. Machine Vision Condition Monitoring of Heavy-Axle Load Rail car Structural Underframe Components, (June).
Seuring, S., 2013. A review of modeling approaches for sustainable supply chain management. Decis. Support Syst. 54 (4), 1513–1520.
Seuring, S., Müller, M., 2008. From a literature review to a conceptual framework for sustainable supply chain management. J. Cleaner Prod. 16 (15), 1699–1710.
Shafiullah, G.M., Ali, A.B.M.S., Thompson, A., Wolfs, P.J., 2010. Predicting vertical acceleration of railway wagons using regression algorithms. IEEE Trans. Intell. Transp. Syst. 11 (2), 290–299.
Shang, H., Berenguer, C., 2014. A Colored Petri Net model for railway track maintenance with two-level inspection. In: European Safety and Reliability Conference (ESREL 2014). Taylor & Francis (CRC Press/Balkema), pp. 1227–1235.
Shao, F., Li, K., Xu, X., 2016. Railway accidents analysis based on the improved algorithm of the maximal information coefficient. Intell. Data Anal. 20 (3), 597–613.
Sharma, S., Cui, Y., He, Q., Li, Z., 2017. Data-driven optimization of railway track maintenance using markov decision process. In: Proceedings of 96th Transportation Research Board Annual Meeting Washington DC.
Silla, A., Kallberg, V.P., 2012. The development of railway safety in Finland. Accid. Anal. Prev. 45, 737–744.
Singh, S., Howard, C.Q., Hansen, C.H., Singh, S., Howard, C.Q., Hansen, C.H., 2015. And Extended Defects Preprint If Accepted for Publication, We Encourage Authors to Link from the Preprint to Researchers have Access to the Formal Publications on ScienceDirect, and So Links Please Note: An Extensive Review of Vibration Modelling of ro, (September), pp. 300–330.
Skarlatos, D., Karakasis, K., Trochidis, A., 2004. Railway wheel fault diagnosis using a fuzzy-logic method. Appl. Acoust. 65 (10), 951–966.
Soleimanmeigouni, I., Ahmadi, A., Kumar, U., 2016. Track geometry degradation and maintenance modelling: a review. Proc. Inst. Mech. Eng., Part F: J. Rail Rapid Transit 232 (1), 73–102.
Starke, P., Walther, F., Eifler, D., 2006. PHYBAL-A new method for lifetime prediction based on strain, temperature and electrical measurements. Int. J. Fatigue 28 (9), 1028–1036.
Stratman, B., Liu, Y., Mahadevan, S., 2007. Structural health monitoring of railroad wheels using wheel impact load detectors. J. Fail. Anal. Prev. 7 (3), 218–225.
Su, Z., Nunez, A., Baldi, S., De Schutter, B., 2016. Model predictive control for rail condition-based maintenance: a multilevel approach. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), vol. 19, pp. 354–359.
Summit, B.D., 2014. Big Data: Challenges in Agriculture, (November), pp. 1–10.
Sun, L. B., Lee, D.H., Erath, A., Huang, X., 2012. Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 142–148.
Sun, L., Lu, Y., Jin, J.G., Lee, D.H., Axhausen, K.W., 2015a. An integrated Bayesian approach for passenger flow assignment in metro networks. Transport. Res. Part C: Emerg. Technol. 52, 116–131.
Sun, Y., Leng, B., Guan, W., 2015b. A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system. Neurocomputing 166, 109–121.
Sura, V.S., 2011. Failure Modeling and Life Prediction of Railroad Wheels. Vanderbilt University.
Tan, K.H., Zhan, Y.Z., Ji, G., Ye, F., Chang, C., 2015. Harvesting big data to enhance supply chain innovation capabilities: an analytic infrastructure based on deduction graph. Int. J. Prod. Econ. 165, 223–233.
Thaduri, A., Galar, D., Kumar, U., 2015. Railway assets: a potential domain for big data analytics. In: Procedia Computer Science, vol. 53, pp. 457–467.
Tipaldo, G., 2014. L'analisi del contenuto ei mass media. Il Mulino, Itinerari.
Tramways and Urban Transits, 2014. Big data, big opportunities. Tramways and Urban Transit Magazine, Nov 17, 2014 (http://www.tautonline.com/big-data-big-opportunities/)
Tsai, T.H., Lee, C.K., Wei, C.H., 2009. Neural network based temporal feature models for short-term railway passenger demand forecasting. Expert Syst. Appl. 36 (2 PART 2), 3728–3736.
Turner, C., Tiwari, A., Starr, A., Blacktop, K., 2016. A review of key planning and scheduling in the rail industry in Europe and UK. Proc. Inst. Mech. Eng. Part F: J. Rail Rapid Transit 230 (3), 984–998.
Tyler Dick, C., Barkan, C., Chapman, E., Stehly, M., 2003. Multivariate statistical model for predicting occurrence and location of broken rails. Transp. Res. Rec. 1825 (1), 48–55.
Vale, C., Bonifácio, C., Seabra, J., Calçada, R., Mazzino, N., Elisa, M., Grimes, D., 2016. Novel efficient technologies in europe for axle bearing condition monitoring - the MAXBE project. Transp. Res. Proc. 14, 635–644.
Van Der Hurk, E., Kroon, L., Maróti, G., Vervest, P., 2015. Deduction of passengers' route choices from smart card data. IEEE Trans. Intell. Transp. Syst. 16 (1), 430–440.
van Gulijk, C., Hughes, P., Figueres-Esteban, M., Dacre, M., Harrison, C., 2015. Big data risk analysis for rail safety? In: Safety and Reliability of Complex Engineered Systems - Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015, pp. 643–650.
Wang, R., Work, D.B., 2015. Data driven approaches for passenger train delay estimation. In: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2015–Octob, pp. 535–540.
Wang, W., He, Q., Cui, Y., Li, Z., 2017. Journal of Transportation Engineering, Part A: Systems Joint Prediction of Remaining Useful Life and Failure Type of Train Wheelsets: A Multi-task Learning Approach.
Wei, S., Yuan, J., Qiu, Y., Luan, X., Han, S., Zhou, W., Xu, C., 2017. Exploring the potential of open big data from ticketing websites to characterize travel patterns within the Chinese high-speed rail system. PLoS ONE 12 (6), 1–13.
White, T., 2015. Hadoop: The Definitive Guide. O'Reilly Media Inc.
Wu, C., Cao, C., Sun, Y., Li, K., 2015. Modeling and analysis of train rear-end collision accidents based on stochastic petri nets. Math. Probl. Eng. 2015.
Yaghini, M., Khoshraftar, M., Seyedabadi, M., 2013. Railway passenger train delay prediction via neural network model. J. Adv. Transport. 47 (3), 355–368.
Yang, C., Letourneau, S., 2005. NRC Publications Archive Archives des publications du CNRC. Chicago, Illinois, USA: The Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005).
Yang, C., Létourneau, S., 2005. Learning to predict train wheel failures. In: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining - KDD '05, (Kdd), pp. 516.
Yang, C., Létourneau, S., 2009. Two-stage classifications for improving time-to-failure estimates: a case study in prognostic of train wheels. Appl. Intell. 31 (3),

255–266.
Yilboga, H., Eker, Ö. F., Güçlü, A., Camci, F., 2010. Failure prediction on railway turnouts using time delay neural networks. In: CIMSA 2010 - IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Proceedings, pp. 134–137.
Yin, J., Zhao, W., 2016. Fault diagnosis network design for vehicle on-board equipments of high-speed railway: a deep learning approach. Eng. Appl. Artif. Intell. 56 (October), 250–259.
Yousefikia, M., Moridpour, S., Setunge, S., Mazloumi, E., 2014. Modeling degradation of tracks for maintenance planning on a tram line. J. Traffic Logist. Eng. 2 (2), 86–91.
Yu, X., Starke, M.R., Tolbert, L.M., Ozpineci, B., 2007. Fuel cell power conditioning for electric power applications : a summary. IET Electr. Power Appl. 1 (5), 643–656.
Zarembski, A.M., 2015. Some examples of big data in railroad engineering. In: Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014, pp. 96–102.
Zarembski, A.M., Einbinder, D., Attoh-Okine, N., 2016. Using multiple adaptive regression to address the impact of track geometry on development of rail defects. Constr. Build. Mater. 127, 546–555.
Zhang, B., Zhang, N., Li, H., Liu, F., Miao, K., 2009. An efficient cloud computing-based architecture for freight system application in China railway. Cloud Comput. 359–368.
Zhang, S., Wang, C., Yang, Z., Chen, Y., Li, J., 2016. Automatic railway power line extraction using mobile laser scanning data. Proc. Int. Arch. Photogram. Rem. Sens. Spat. Inform. Sci. Prague, Czech Republic 5, 615–619.
Zhang, Z., He, Q., Zhu, S., 2017. Potentials of using social media to infer the longitudinal travel behavior: a sequential model-based clustering method. Transport. Res. Part C: Emerg. Technol. 85 (February), 396–414.
Zhao, L.-H., Zhang, C.-L., Qiu, K.-M., Li, Q., 2013. A fault diagnosis method for the tuning area of jointless track circuits based on a neural network. Proc. Inst. Mech. Eng., Part F: J. Rail Rapid Transit 227 (4), 333–343.
Zhao, Y., Xu, T., Hai-feng, W., 2014. Text mining based fault diagnosis of vehicle on-board equipment for high speed railway. In: 2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 900–905.
Zhi, S., Li, J., Zarembski, A.M., 2016. Predictive modeling of the rail grinding process using a distributed cutting grain approach. Proc. Inst. Mech. Eng. Part F: J. Rail Rapid Transit 230 (6), 1540–1560.
Zhu, M., Cheng, X., Miao, L., Sun, X., Wang, S., 2013. Advanced stochastic modeling of railway track irregularities. Adv. Mech. Eng. 2013.
Zhang, Z., He, Q., Gao, J., Ni, M., 2018. A deep learning approach for detecting traffic accidents from social media data. Transport. Res. Part C: Emerg. Technol. 86 (December 2017), 580–596.
Zilko, A.A., Kurowicka, D., Goverde, R.M.P., 2016. Modeling railway disruption lengths with Copula Bayesian Networks. Transport. Res. Part C: Emerg. Technol. 68, 350–368.