University of Huddersfield Repository

EL Rashidy, Rawia Ahmed Hassan, Hughes, Peter, Figueres-Esteban, Miguel, Harrison, Chris and Van Gulijk, Coen

A Big Data modeling approach with graph databases for SPAD risk

Original Citation

EL Rashidy, Rawia Ahmed Hassan, Hughes, Peter, Figueres-Esteban, Miguel, Harrison, Chris and Van Gulijk, Coen (2017) A Big Data modeling approach with graph databases for SPAD risk. Safety Science. ISSN 0925-7535

This version is available at http://eprints.hud.ac.uk/id/eprint/33955/

The University Repository is a digital collection of the research output of the University, available on Open Access. Copyright and Moral Rights for the items on this site are retained by the individual author and/or other copyright owners. Users may access full items free of charge; copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational or not-for-profit purposes without prior permission or charge, provided:

- The authors, title and full bibliographic details is credited in any copy;
- A hyperlink and/or URL is included for the original metadata page; and
- The content is not changed in any way.

For more information, including our policy and submission procedure, please contact the Repository Team at: E.mailbox@hud.ac.uk.

http://eprints.hud.ac.uk/

A Big Data modeling approach with graph databases for SPAD risk

R.A.H. EL Rashidy[1], P. Hughes[1], M. Figueres-Esteban[1], C. Harrison[2], C. Van Gulijk[1].

[1]*Institute of Railway Research, University of Huddersfield, Huddersfield, UK*
[2]*RSSB, London UK*

Abstract

This paper proposes a model to assess train passing a red signal without authorization, a SPAD. The approach is based on Big Data techniques so that many types of data may be integrated, or even added at a later date, to get a richer view of these complicated events. The proposed approach integrates multiple data sources using a graph database. A four-steps data modeling approach for safety data model is introduced. The steps are problem formulation, identification of data points, identification of relations and calculation of the safety indicators. A graph database was used to store, manage and query the data, whereas R software was used to automate the data upload and post-process the results. A case study demonstrates how indicators have extracted that warning in the case that the SPAD safety envelope is reduced. The technique is demonstrated with a case study that focuses on the detection of SPADs and safety distances for SPADs. The latter provides indicators for to assess the severity of near-SPAD incidents.

1 Introduction

Railway systems create an incredible amount of data that potentially hold safety learning if it can be tapped into. The GB railways are exploring several ways to unlock safety learning (Network Rail, 2014; RSSB, 2016). These efforts benefit from the development of new technologies to deal with such data. This paper focuses on using graph databases which, are particularly useful for safety analysis and management with big data sources (Hoffer et al., 2016; Sadalage and Fowler, 2013). A key feature is that various data sources can be stored alongside one another in the same database to create a more detailed understanding of safety than is possible by considering each source separately. This bypasses the need for computationally expensive join operations for traditional SQL (Structured Query Language) data tables (Miller, 2013).

This paper introduces methods to combine railway data sources in scalable graph databases to improve the understanding of the underlying factors of signals passed at danger (SPAD). The approach focuses on recognizing SPAD-related safety occurrences in large amounts of data that were, initially, not designed to detect SPADs. The approach provides a means to detect near-miss aspects of SPAD risk that have not previously been understood which potentially feed into driver behavior management.

2 Background

2.1 SPADs

SPADs are events where a train passes a stop signal and proceeds onto a section of track where it does not have authority. SPADs can lead to trains colliding with other trains or road vehicles on level crossings, derailing, or striking workers and equipment. As such, SPADs present a major safety risk to the railway. In Britain, a SPAD at Ladbroke Grove resulted in 31 fatalities in 1999 (Health and Safety Executive, 2000; Lawton and Ward, 2005; Stanton and Walker, 2011). Since then, considerable efforts have been made by the rail industry to reduce the number of SPADs.

With the systems currently used on the railway, understanding of the underlying causes of SPADs comes largely from analysis by safety experts after a SPAD has occurred. Analysis reports, such as Rail Accident Investigation Branch (2016a; 16b) provide examples of such analyses and provide recommendations to prevent recurrence of similar incidents. Whilst such a retrospective approach is clearly meaningful, it is an aspiration of railway safety staff to be able to identify the causes of accidents prior to the accident occurring.

Nikandros and Tombs (2007) addressed the issue by taking a data-centric approach that takes input from the train control systems and allows SPADs to be normalized by the number of times trains approach stop signals. This approach allows not only an understanding of the number of SPADs that have occurred but also the number of times train drivers successfully stop the train before a stop signal. Zhao et al. (2016) extended this approach by analyzing several years' worth of Train Describer (TD) data that was downloaded from the TD-live data stream provided by Network Rail. This was a step towards big-data techniques since it consists of almost two billion records in a single year.

The approach by Nikandros and Tombs (2007) and Zhao et al. (2016) provides useful insights but additional data sources could enrich the insights further. Green et al. (2011) described a method of using data collected from the On Train Data Recording (OTDR) equipment to assess driver performance. Since driver performance is a significant contributor to SPADs (Dhillon, 2007); a number of studies, for example, Naweed (2013), Gibson et al. (2007), Kyriakidis et al. (2015) and Wright et al. (2007) sought to identify the factors that influence human behavior that could contribute to SPADs. This paper progresses beyond the opportunity to extend the work of Nikandros and Tombs (2007), Zhao et al. (2016), Green et al. (2011) and Dhillon (2007). The aim is to create an understanding of SPADs not only from the state of the signaling and the number of times trains approach stop signals but also from the performance of train drivers on the approach to signals. This paper describes an efficient data analysis approach for combining this data but, for reasons of confidentiality, cannot present real data.

2.2 Data management and analysis

Traditional relational databases (or SQL databases) have proven to be effective for relatively "small" amounts of data due to their speed and due to unimpeded data access. The key to SQL success is that they use a relational table that remembers where data are stored exactly and which type of data it is. The relational table enables ACID (atomicity, consistency, isolation, and durability). However, high volume datasets and the complex data structures make SQL databases unwieldy and difficult to write queries for (Cudré-Mauroux and Elnikety, 2011). A solution to bypass such problems is to omit the relational table by simply storing data in a system that, for lack of a better example, finds its analogy in an infinitely scalable library card catalog. In a library catalog, numerous pieces of information are stored on cards with label indexes in labeled boxes. Only a very basic index, usually alphabetic, states the approximate location of cards but the system does not drill down to each exact card. Retrieving that information requires a query for that information and a person and/or search engine to find the

relevant card. “Not only SQL databases” (aka NoSQL databases) work in that way. In addition to that, the data does not have to be of any specific format or stored in any particular order. Numerous pieces of digital information (often files of arbitrary type) are stored under a ‘key-value’ index label (identifying a unique information file), in an almost infinitely scalable database (Sadalage and Fowler, 2013).

For this work, a particular type of NoSQL database is used: a graph database in which a graph overlays the data to create connections. The fundamental units of a graph database are nodes and edges. Unique labels identify the type and content of the nodes and edges (e.g., Signal, Train). In this way, it is possible to develop flexible data models that support data demands in complex domains such as medicine, biology, chemistry and social networking (Angles et al., 2013; Jouili and Vansteenberghe, 2013). Moreover, graph databases provide visual interfaces that enable users to perceive their data whilst they are performing their data analysis. Graphs enhance discovery from data which is useful for analysis (Figueres-Esteban et al., 2016a; Figueres-Esteban et al., 2016b; Miller, 2013). For safety analysis, graph databases provide a flexible platform in the sense that the analyst can introduce additional data for his/her risk problem, even if it is of a different type (numeric, visual or text). This paper explains how the technology is adopted for SPADs.

3 Method

The modeling approach comprises of four steps that are described successively in the following paragraphs. Three different data sources are used for the SPAD safety data model; *viz.* TD data, OTDR data and signaling location data. The Neo4J software was used to store, manage and query the data. R software was used to automate the data upload and post-process the results. The method is described in details below.

3.1 Safety Data model

The first step is the development of a safety data model. Four key steps were used to construct the safety model, *viz.* the problem formulation, the identification of data points (which are represented as nodes), the identification of relationships between the nodes (the edges) and the required indicators. Figure 1 shows the four elements of the safety data model for this particular investigation; they are described in some more detail below.

The first part in the development of the safety data model is the problem formulation. For this paper, it was derived from a narrowly defined research question:

- Which train service stopped at a red aspect and/or had a SPAD?

This informs the second part, the definition and content of relevant data points that are required for the data model:

- Service node; defined as the complete service from origin to destination (including data about driver number, vehicle number, and start-time).
- Service-instance node; defined as a single data-row in the OTDR file (including data about time, location and speed, amongst others)
- Signal node; defined as signal as found in the TD feed (including data about signal ID and location),
- Red aspect node; defined by a red-aspect approach algorithm (including data about signal ID, starting time and ending time)

These nodes were created in the graph database by Cypher queries, as shown in Table 1.

The relations make up the third part. The relations are represented by edges in the graph. In the case presented in this paper, three relations were relevant:

- The relationship between the service node and service-instance node,
- The relationship between the signal node and red aspect node, and
- The relationship between service-instance node and the red aspect node.

These edges were created by queries, as shown in Table 1.

The safety indicators represent the final part of the data model. They are queries that filter the database to provide insight into the research question from step 1. In this investigation, the following safety indicators were used:

- Number of SPADs per signal;
- Number of red aspects approached by a service.

The indicators are extracted by queries; examples are given in Section 3.4.

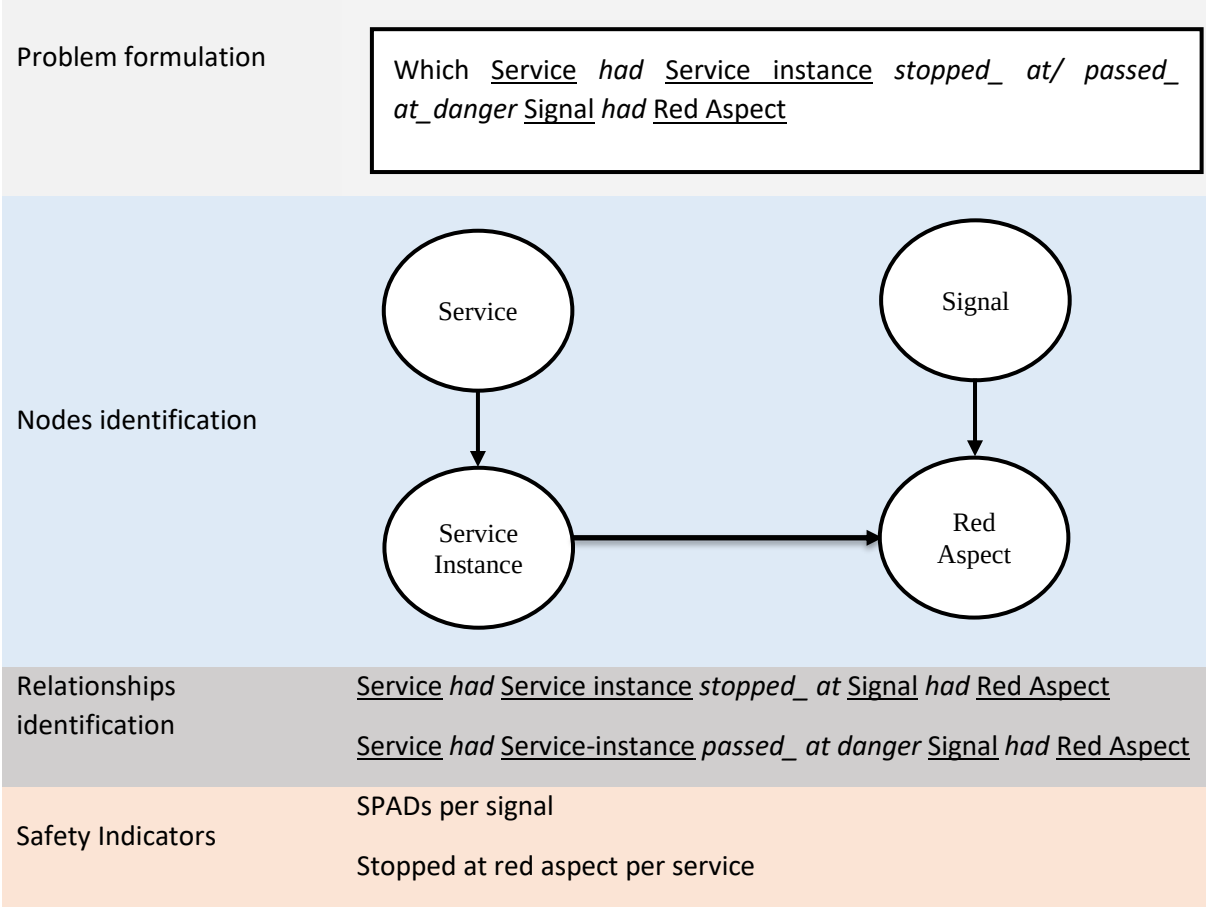


Figure 1: Four steps of a safety data model.

3.2 Data sources

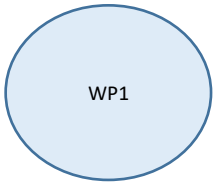
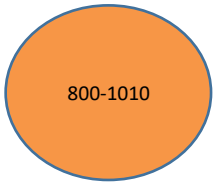
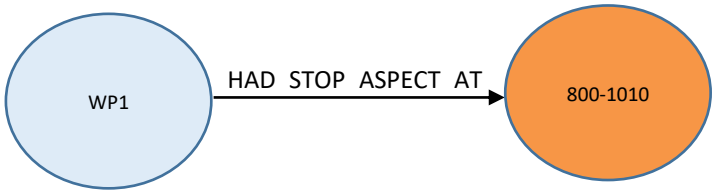
The identification of data sources, the second step in this method, and the design of the safety data model typically take place at the same time. However, there is a danger in first identifying data and formulating a research question later; it can lead to research bias in the sense that relevant research questions might be adjusted or simplified to match the data. In this work, both steps were performed

simultaneously. Due to data-sensitivity, this investigation used synthetic data replicated from three data sources: OTDR (On Train Data Recorder), TD-data (Train Describer) and Signal location data. The use of artificial data also adds to the clarity of this paper as significant efforts for data-parsing and cleansing pose distractions for explaining the development of a safety data model. The case study treats a small section of railway covering four signals and four services running along the signals. The data included four services, 3B01, 3B02, 3B03 and 3B04, and four signals named WP1, WP2, WP3, and WP4. Time, speed, and location were simulated for each service, whereas location and red aspect time were given per signal.

3.3 Data handling

Data handling, the third step, builds the actual safety model based on datasets in the database. Table 1 illustrates the creation of nodes and relations using Cypher language for queries. Cypher is considered as a declarative query language as it focuses on the aspects of the result rather than on methods to obtain the result. Examples are given in Table 1 (for more details about Cypher see Panzarino, 2014)).

Table 1: Cypher Query Examples.

Query	database instance	
CREATE(n:ServiceInstance {name: '3B04-2694', time:2694, location:21669.14, speed:21.26, type: 'Service_Instance'}) RETURN n	name 3B04-2694 time 2694 location 21669.14 speed 21.26 type Service_Instance	
CREATE (signal:SIGNAL{name: 'WP1',type:'SIGNAL'})	name WP1 type SIGNAL	
CREATE (redAspect:Red_Aspect{name: '800-1010', Signal: 'WP1', startTime: 800, endTime: 1010, signalLocation:3700,type: 'Red_Aspect'})	signalLocation 3700 name 800-1010 signal WP1 startTime 800 endTime 1010 type Red_Aspect	
MATCH (n: SIGNAL {name: 'WP1'}), (m : Red_Aspect { signal: "WP1" }) MERGE (n) [rel:HAD_STOP_ASPECT_AT]->(m) RETURN n, m		

3.4 Data analysis

The last step is the data analysis to identify the safety indicators in the data. The relationships are identified by identifying data nodes where several conditions are met simultaneously. The relationships were identified based on the synchronization between the OTMR, TD feed and signal data. The conditions presented in Table 2 were coded in Cypher in a similar way as in Table 1.

Table 2: Relations for indicators.

Relation	Conditions that identify the Relation
SPAD (Indicator 1: SPADs per signal)	Service-instance location(t_1)>Signal had a red aspect (t_{start},t_{end}) location<Service-instance location(t_2) Service-instance speed(t_1) >0 $t_{start}<t_1<t_{end}$ $t_2<t_{end}$
Service-instance stopped at a red aspect (Indicator 2: Stopped at red aspect per service)	Service-instance location(t_1) > Signal had a red aspect (t_{start},t_{end}) location < Service-instance location(t_2) Service-instance speed(t_1) = 0 $t_{start}<t_1<t_{end}$ $t_2>t_{end}$
services-instance approaching a red aspect (Indicator 2: Stopped at red aspect per service)	Service-instance location(t_1)>Signal had a red aspect (t_{start},t_{end}) location<Service-instance location(t_2) Service-instance speed(t_1) >0 $t_1>t_{end}$ $t_2>t_{end}$

where:
t_1 is the timestamp (in seconds) of the last service instance in front of a particular signal location,
t_2 is the timestamp (in seconds) of the first service instance behind a particular signal location,
t_{start} is the start time; the second that the particular signal changes to a red aspect,
t_{end} is the end time; the second that the particular signal clears from a red aspect.

4 Results

Table 3 summarizes SPAD events for signals. The red aspect start time, end time (in seconds relative to the starting time of the train service), the service, and its travel speed at SPAD are given in Table 3. The three signals that had SPADs were investigated further to check how many services had to stop by these three signals. It should be noted that the data used to illustrate the method is simulated data where SPAD had occurred due to simulated over speed profile; in reality, SPADs tend to be rare events.

Table 3: SPAD per signal.

Signal	Signal at Red		Time of SPAD	Travel speed at SPAD	Train Service
	Start time t_{start} (s)	End time t_{end} (s)	t_2 (s)	(m/s)	(Journey number)
WP1	800	1010	898	16	3B03
WP3	1932	2059	2013	1	3B04
WP4	1243	1550	1469	21	3B01
WP4	1942	2093	2055	21	3B03

The analysis was also carried out at the service level, Table 4 presents all information related to Service 3B01, i.e. the service had stopped twice and had one SPAD.

Table 4: Stopped/SPAD at red aspect per service for 3B01.

Signal	Signal at Red		Time of SPAD	Travel speed at SPAD	Train service Status at signal
	Start time t_{start} (s)	End time t_{end} (s)	t_2 (s)	(m/s)	
WP1	0	230	209	0	Stopped
WP2	551	799	774	0	Stopped
WP4	1243	1550	1469	21	SPAD

5 Analysis and discussion

The technique proposed in this paper demonstrates a hands-on safety data modeling problem. It shows that it is technically feasible to combine operational data from different sources to identify safety issues. The case study demonstrates not only which signals and services have had SPADs but also how many red aspect approaches a single service encounters and at what speed red aspects are approached or passed.

The techniques allows better insight in SPADs and helps the decision makers develop safety indicators to monitor and investigate the link between near-SPADs and actual SPADs.

The graph database easily facilitates extensions of the analysis to include the braking behavior of the train whilst it had a SPAD and the distance to signal where it stopped before a red aspect by extending the query. Table 5 shows just how flexible the graph approach is: straightforward analytics can be added to the data model to estimate distances to stop based on data that is already in the database. Table 5 gives an example where three services were could stop safely using the emergency brake far away from the signal.This provides the basis for an automated safety indicator for SPADs without the need for an in-depth investigation by experts which, potentially, could be completely automated. The traveled speed of Service 3B02 was significantly higher than the other two services, which means it would score higher on a risk scale.

It should be noted that the safety margin needed may be affected by the track conditions such as low adhesive condition. The simulated data used in this study did not consider the variations in the track conditions that may lead to different scenarios such as low adhesive conditions. A number of braking behavior simulations such as the one developed by Meli et al. (2014) and Pugi et al. (2013) could be used to take into account the impact of track condition on the safety margin. Another approach is to 'train' the data-model with recorded, real-life approaches to that same signal.

Table 5: Additional analysis on signal cleared prior to a service approach.

Signal	Signal at Red		Train Service (Journey number)	Travel speed (m/s)	Stop distance between train and signal (m)	Emergency braking distance @12%g (m)	Safety margin (m)
	Start time t_{start} (s)	End time t_{end} (s)					
WP3	790	986	3B01	11	195	46	149
WP3	1011	1286	3B02	22	254	202	52
WP3	1498	1563	3B03	12	241	57	184

Using the safety data model approach with graph databases, it is relatively straightforward to extend the analysis to incorporate the effects of factors such as weather conditions, wheel adhesion, and service disruption to assess the safety state and safety indicators. If additional data is required, the four steps for the safety data model simply have to be repeated to identify additional research questions, nodes, relationships and indicators. The technique described in this paper lends itself to be indefinitely scaled to include additional research questions. The graph database adds the flexibility to deal with these multiple research questions, and additional data sources. In that way, this technique paves the way toward achieving one of the aspirations of safety management: proactive safety interventions that can be demonstrated to have reduced risk even before any accidents occur.

A particular application area for this technique is with the signal overrun risk assessment tool (SORAT) that assesses the SPAD risk at signals. The UK Network Rail has a rolling five-year program whereby all signals have their SPAD risk assessed. Some are looked at in more detail, but it is a largely manual process of entering data and running it through a model. The advantage that the approach described in this paper offers is the opportunity to automate some of this analysis and introduce new metrics that can better aid the understanding of SPAD risk at signals and their underlying causes. Furthermore, the approach described in this paper provides an opportunity to identify underlying causal factors that may otherwise not be detected.

6 Conclusion

This paper demonstrates a Big Data modeling approach for safety based on graph databases. It addresses SPAD risk as a case study. The key is a consistent approach to building a safety data model integrating multiple sources of data. This paper offers a straightforward method to provide such consistency.

The results in the case study are limited in the number of data sources used, and in the safety information that has been provided. However, the technique demonstrates the basis that can be extended for additional data sources, and to uncover additional factors that may affect SPAD risk on the railway. From the exercises in the case study we infer that SPAD risk can be understood in new ways by the application of new data sources, for example the method can provide a new understanding of human factors and driver behavior that affect the risk. Hypothetically further data sources such as localized weather conditions including sun angle, or even factors such as timetable data and train on-time running data could be included to broaden our understanding. Such new sources of data provide the ability to analyze the complexities of SPAD risk to be understood in ways that have not previously been possible.

In more general terms, the flexibility embedding a safety data model in a graph database makes it useful in practically every safety and risk domain. Considering the potential for scaling graph databases to extremely large data sets, and developing complex queries, it is not yet clear what limits there are to extending the approach.

Acknowledgements

This work was funded by the Rail Safety and Standards Board (RSSB).

References

Angles, R., Prat-Pérez, A., Dominguez-Sal, D., Larriba-Pey, J. L., 2013. Benchmarking Database Systems for Social Network Applications, in: First International Workshop on Graph Data Management Experiences and Systems, GRADES ’13. ACM, New York, NY, USA, pp. 1:7.

Cudré-Mauroux, P., Elnikety, S., 2011. Graph data management systems for new application domains. Proc. VLDB Endow. 4.

Dhillon, B.S., 2007. Human reliability and error in transportation systems. Springer Verlag, London.

European Commission, 2004. Directive 2004/49/EC, L.220/16. Off. J. Eur. Union.

Figueres-Esteban, M., Hughes, P., Van Gulijk, C., 2016a. Visual analytics for text-based railway incident reports. Saf. Sci. 89, pp. 72:76.

Figueres-Esteban, M., Hughes, P., Van Gulijk, C., 2016b. Using visual analytics to make sense of railway

Close Calls. Proc IMechE Part F J. Rail Rapid Transit 0, pp. 1:8.

Gibson, W.H., Shelton, J., Mills, A., 2007. The Impact of returning from rest days on SPAD incidents, in: Wilson, J.R., Norris, B., Clarke, T., Mills, A. (Eds.), People and Rail Systems. Ashgate Publishing Limited, London, pp. 475:481.

Green, S.R., Barkby, S., Puttock, A., Craggs, R., 2011. Automatically assessing driver performance using black box OTDR data. Proc. 5th IET Conf. Railw. Cond. Monit. Non-Destructive Test. pp. 22:22.

Health and Safety Executive, 2000. The Ladbroke Grove Rail Inquiry, Part 1 - Report. HSE BOOKS, Norwich.

Hoffer, J., Venkataraman, R., Topi, H., 2016. Modern Database Management. Pearson Education, Inc, New Jersey.

Jouili, S., Vansteenberghe, V., 2013. An Empirical Comparison of Graph Databases, in: 2013 International Conference on Social Computing. pp. 708:715.

Junghanns, M., Kießling, M., and Averbuch, A., 2017. Cypher-based Graph Pattern Matching in Gradoop, GRADES'17, Chicago, IL, USA

Kyriakidis, M., Majumdar, A., Ochieng, W.Y., 2015. Data based framework to identify the most significant performance shaping factors in railway operations. Saf. Sci. 78, pp. 60:76.

Lawton, R., Ward, N.J., 2005. A systems analysis of the Ladbroke Grove rail crash. Accid. Anal. Prev. 37, pp. 235:244.

Meli, E., Pugi, L., Ridolfi, A. 2014. An innovative degraded adhesion model for multibody applications in the railway field Multibody System Dynamics, 32 (2), pp. 133-157.

Miller, J.J., 2013. Graph database applications and concepts with Neo4j, in: Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA. p. 36.

Naweed, A., 2013. Psychological factors for driver distraction and inattention in the Australian and New Zealand rail industry. Accid. Anal. Prev. 60, pp. 193:204.

Network Rail, 2014. Asset management strategy. London.

Nikandros, G., Tombs, D., 2007. Measuring railway signals passed at danger, in: Proceedings 12th Australian Conference on Safety-Related Programmable Systems. Australian Computer Society, Inc., Adelaide, pp. 41:46.

Panzarino, O., 2014. Learning Cypher. Packt Publishing, Birmingham, UK.

Pugi, L., Malvezzi, M., Papini, S., Vettori, G. 2013. Design and preliminary validation of a tool for the simulation of train braking performance Luca Pugi Monica Malvezzi Susanna Papini, Journal of Modern Transportation, 21 (4), pp. 247-257.

RAIB, Rail Accident Investigation Branch), 2016a. Two signal passed at danger incidents, at Reading Westbury Line Junction, 28 March 2015, and Ruscombe Junction, 3 November 2015. Rail Accid. Rep.

RAIB, Rail Accident Investigation Branch, 2016b. Signal passed at danger on approach to Wootton Bassett Junction, Wiltshire 7 March 2015. Rail Accid. Rep.

Railway Group UK, 2003. Railway Group Safety, Performance Monitoring – Definitions and Guidance, Issue 2, GE/GN8510tle.

RSSB, 2016. SMIS + Developing the new Safety Management Intelligence System.

Sadalage, P.J., Fowler, M., 2013. NoSQL Distilled: A brief guide to the emerging word of polyglot persistence. Pearson Education, Inc, New Jersey.

Stanton, N.A., Walker, G.H., 2011. Exploring the psychological factors involved in the Ladbroke Grove rail accident. Accid. Anal. Prev. 43, pp. 1117:1127.

Van Gulijk, C., Dennis, C., 2016. Big Data Risk Analysis – linking wider business and safety information systems for improved safety management. Saf. Reliab. 36, pp. 131:133.

Van Gulijk, C., Hughes, P., Figueres-Esteban, M., Dacre, M., Harrison, C., 2015. Big Data Risk Analysis for Rail Safety?, in: Podofillini, L., Sudret, B., Stojadinovic, B., Zio, E., Kroger, W. (Eds.), Safety and Reliability of Complex Engineered Systems. Taylor & Francis Group, London, pp. 643:650.

Wright, L., Dabekaussen, M., Van der Schaaf, T., 2007. Predicting the causes of Spad Incidents, in: Wilson, J.R., Norris, B., Clarke, T., Mills, A. (Eds.), People and Rail Systems. Ashgate Publishing Limited, London, pp. 491–497.

Zhao, Y., Stow, J., Harrison, C., 2016. Estimating the frequency of trains approaching red signals: a case study for improving the understanding of SPAD risk. IET Intell. Transp. Syst. 10(7), pp. 579:586.