Proceedings of the 7th International Young Geotechnical Engineers Conference – Scott (Ed.)

Investigating railway track load bearing capacity using descriptive data mining

Étude de la capacité de charge des voies ferrées à l'aide de l'exploration de données descriptives

Mikko Sauni
Research Centre Terra, Tampere University, Finland, mikko.sauni@tuni.fi

ABSTRACT: The load bearing capacity is difficult to determine throughout a railway track section in practice, as it is generally unfeasible to provide investigations and calculations for all the structure variations on an entire section. Nevertheless, there is plenty of data concerning track structure features and the observed resilience of track sections. This provides an opportunity to investigate the correlations between combinations of structure features and load bearing capacity anomalies, which have actualised as track geometry deterioration. This study investigated whether a descriptive data mining method Generalized Unary Hypothesis Automata (GUHA) could be utilised for that purpose. The data set used in this study consisted of the track geometry measurement history, GPR measurements, LiDAR point clouds, track deflection measurements, and asset data. The GUHA method provided hypotheses about the track structure features associated with rapid track geometry deterioration. The results of GUHA were found useful in identifying actual phenomena governing the track structure behaviour from the railway data. Thus, GUHA provides a practical tool for investigating the features influencing the load bearing capacity on a track section scale.

RÉSUMÉ: La capacité portante est difficile à déterminer sur l'ensemble d'une section de voie ferrée dans la pratique, car il est généralement impossible de fournir des études et des calculs pour toutes les variations de structure sur l'ensemble de la section. Néanmoins, il existe de nombreuses données concernant les caractéristiques de la structure de la voie et la résilience observée des sections de voie. Cela donne l'occasion d'étudier les corrélations entre les combinaisons de caractéristiques de la structure et les anomalies de capacité de charge, qui se sont traduites par une détérioration de la géométrie de la voie. Cette étude a examiné si une méthode d'exploration de données descriptives Generalized Unary Hypothesis Automata (GUHA) pouvait être utilisée à cette fin. L'ensemble de données utilisé dans cette étude se composait de l'historique des mesures de la géométrie de la voie et GPR, des nuages de points LiDAR, des mesures de déflexion de la voie et des données sur les actifs. La méthode GUHA a fourni des hypothèses sur les caractéristiques de la structure de la voie associées à la détérioration rapide de la géométrie de la voie. Les résultats de GUHA se sont avérés utiles pour identifier les phénomènes réels régissant le comportement de la structure de la voie à partir des données ferroviaires. Ainsi, GUHA fournit un outil pratique pour étudier les caractéristiques influençant la capacité de charge sur une échelle de section de voie.

KEYWORDS: data mining, load bearing capacity, railway, track structure

1 INTRODUCTION

Track load bearing capacity can be described as the ability of a railway structure to resist permanent deformations due to repeated loading. The load bearing capacity can be exceeded if the loading is increased or the track structure strength is impaired. In either of these cases, the subsequent result is the deterioration of track geometry, which, consequently, causes maintenance needs and safety concerns and should be minimised. Generally, railway lines' track geometry is monitored periodically using a track measurement car, which provides information about the deviations from an ideal track geometry. Using this information, the locations and severities of problematic areas can be assessed. However, the practical problem with track geometry irregularities is not locating them but assessing the root causes contributing to them. If the root causes of track geometry irregularities are not known, the assigned maintenance might not correct the issue, only temporarily treat the symptom. This causes unnecessary maintenance costs and track down time.

There can be several different causes to track geometry irregularities, and assessing their possible influences requires examining multiple data sources. For example, the subgrade conditions, track structure formation and moisture, and special track structures such as turnouts must all be considered in the assessments. Subjective analysis of all these data sources may be adequate for inspecting a single problematic area, but for inspecting the correlations of different structure types on a track section scale requires computational assistance, as the amount of data is vast and correlations may be difficult to observe visually.

Many previous studies have investigated track load bearing capacity using computational approaches, often modelling track geometry deterioration based on track structure parameters (Higgins & Liu 2018, Soleimanmeigouni et al. 2018). However, before modelling track geometry deterioration, it is pertinent to investigate how different parameters in the available initial data correlate. Despite this, exploratory analysis of actual track structure data is rarely mentioned in previous research.

In this study, a descriptive data mining method, Generalized Unary Hypothesis Automata (GUHA), was used to investigate the causes of realised insufficient track load bearing capacity from a large multivariate dataset. The method was tested on real world data from a track section located in South-Eastern Finland. The aim was to assess the GUHA method's ability to produce practical results from actual railway structure data.

The rest of this article is organised as follows. First, the available data is described. Second, the GUHA method and its use in the railway domain is elaborated. Finally, the results from the data mining are discussed and conclusions are presented.

2 THE INITIAL DATA

The initial data comprised of multiple different data sources, which are presented in Table 1 along with the data type. In Table 1, ratio scale refers to floating point numbers, binary refers to data containing only 0s or 1s, and categorical refers to data consisting of different groups with no particular order, such as clay, sand, and rock.

The track geometry deterioration rate (TGDR) was calculated from a ten-year track geometry history, which contained the longitudinal level (LL) irregularities. Each measurement was calculated using a 20 m running roughness (R2) value (Li et al. 2016). The difference between two consecutive measurements' LL R2 was used to describe the rate of deterioration. If

subsequent measurements' LL R2 significantly decreased, the decrease was considered to have been caused by maintenance of tracks, i.e., tamping, and the decreased values were ignored. The mean of all increased LL R2 values in a particular location was used to describe the TGDR of that location.

Table 1. Initial data sources and data types

Track geometry measurement history	Ratio scale data
Continuous track deflection measurements	Ratio scale data
Track geometry elements	Binary data
GPR measurement data	Ratio scale data
Continuous LiDAR data	Ratio scale data
Sub soil data	Categorical data
Foundation type data	Categorical data
Asset data	Categorical data
Tamping records	Categorical data

Track deflection was measured continuously using a specific measurement device, Stiffmaster, which was developed at Research Centre Terra (Luomala et al. 2017). Track deflection 20 m moving average and variance were used in the data mining. The moving average indicated the amount of deflection, whereas variance indicated changes in deflection. Furthermore, the track deflection measurements provided cant data, which was used to indicate two types of track geometry elements: straights and curves.

GPR measurement data was used in assessing track structural layer thicknesses as well as different structural layers' relative moisture damage indices (MDI). The MDI indicated how much the layer contained fine materials and water compared to the average MDI of the whole track section.

Continuous LiDAR data was used to calculate the drainage level of the embankment. The calculation was based on evaluating the lowest level within a 20 m range along the track for both sides of the track separately. This calculation was used to alleviate the effects of foliage on the measurements.

Other parameters included manual assessment of data from several data bases. Categorical subsoil data was analysed using available soil maps. The foundation type was assessed using video of the track, LiDAR data, and GPR data. Asset data was evaluated and located using video of the track and asset management data bases. Asset data included turnouts, culverts, bridges, and signalling equipment installed in the track superstructure. Transitions zones were assigned to areas, where the track structure changed significantly, e.g., bridge approaches and turnouts. Also, partial tamping records were available and used to indicate the number of past tamping actions.

All this data was formed into a single matrix, in which one row depicted a one metre long section of track described by the columns that represented the different data sources. Multiple data sets from different track sections were tested, but the example provided in this paper concerns a track section, Kouvola–Kotka, located in the South-Eastern coastal area of Finland. This section is 53 km in length, has a maximum speed of 140 km/h for passenger trains, and a maximum axle weight of 22,5 tons for freight trains. The track section was originally built over 50 years ago and exhibits some problematic areas.

3 THE GUHA DATA MINING METHOD

GUHA can be classified as an unsupervised data mining method, meaning there is no training or testing data set, because the objective of data mining is only to explore given input data. GUHA is based on a logic formalism, in which the output of data mining is either true or false. For an output to be true, the input data must support it and *vice versa* (Hájek and Havránek 1978). For practical use, GUHA is implemented in a computer program LISp-Miner. LISp-Miner includes several different modules for different types of data mining tasks. The modules used in this study were 4ft-Miner and AC4ft-Miner.

The generic process of the GUHA method using LISp-Miner is presented in Figure 1. The process begins with forming analytical questions about a subject and gathering relevant data for answering the questions. The initial data is conformed into a matrix, as described in Section 2. Then, the input data is imported, and the LISp-Miner is used to translate the questions into GUHA language. After this, the GUHA data mining is run and hypotheses are produced as results. The user can select a meaningful hypothesis for closer evaluation. The closer evaluation is conducted by interpreting the contingency table of the hypothesis. Finally, if the hypothesis is meaningful and the contingency table indicates that there is sufficient data supporting the hypothesis, the user can translate the hypothesis into a human language and present it. If the produced hypotheses aren't satisfying, the user can tune the data mining quantifiers to adjust the types of results that will be produced.

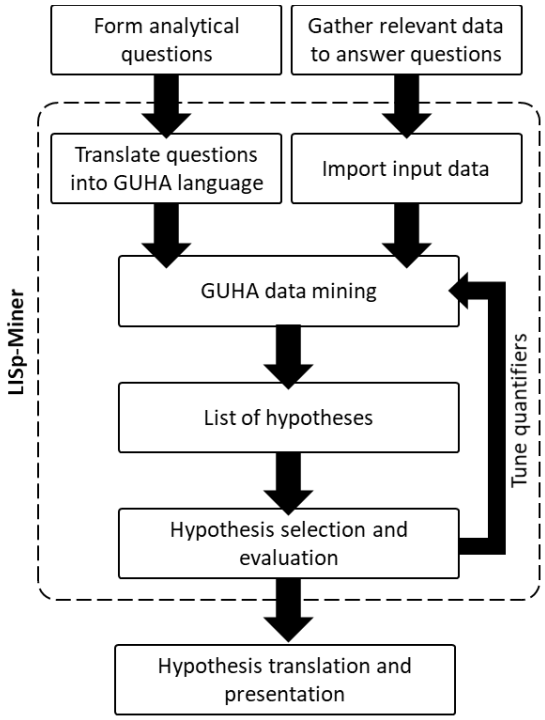

Figure 1. Generic GUHA process when using LISp-Miner.

The translation of the analytical questions into GUHA language is based on using antecedents, succedents, conditions, and quantifiers. Antecedents, succedents, and conditions are certain groups within the initial data. The user creates these groups by dividing each initial data source into categories based on given data mining parameters.

The quantifiers are rules, which the search for hypotheses is based on. Quantifiers work by assessing the contingency table (Table 2) of a potential hypothesis by evaluating the antecedents (ϕ) relationship to the succedent (ψ) when a condition (γ) is satisfied. In the contingency table, the number of objects:

- satisfying both ψ and ϕ is a
- satisfying only ϕ is b
- satisfying only ϕ is c
- not satisfying ψ nor ϕ is d.

Table 2. Contingency table parameters (Berka 2016).

γ	ψ	$\neg\psi$	Σ
ϕ	a	b	$a + b = r$
$\neg\phi$	c	d	$c + d = s$
Σ	$a + c = k$	$b + d = l$	n

In practice, the antecedent (ϕ) denotes the number of observations that are in accordance with the hypothesis precondition parameters. The succedent (ψ) denotes the number of observations in accordance with the parameter that the preconditions are correlated with. The sign ¬ denotes their negation, or, in other words, their opposite. A condition (γ) can be used to limit the *Base* data by assigning rules that ignore certain types of cases, for example if the user wishes to ignore bridges from the data, they can select the condition to remove bridges from the analysis. As an example of quantifiers, if the user is interested in finding the combination of antecedents that have the strongest relationship to a certain succedent, the user can use the founded implication quantifier. The founded implication quantifier assesses contingency table parameters a and b relationship commonness to p, when a frequency is more than the *Base*, by using the following equation (1):

$$\frac{a}{a+b} \geq p \wedge a \geq Base \quad \text{(Rauch 2013).} \tag{1}$$

In practical terms, the relationship between the rows of data that support the hypothesis a and the sum of rows supporting and opposing the hypothesis $a + b$ must be stronger than the defined confidence p. By adjusting the *Base* quantifier, the user can choose how large a data set must support a hypothesis for it to be accepted as a result. A rule of thumb for adjusting the *Base* quantifier is to start with a large *Base* and obtain little to none results. Then the *Base* can be decreased until a reasonable number of hypotheses is obtained. This way, the calculation times are kept down, and the number of hypotheses is controllable. As a generalisation, a reasonable number of obtained hypotheses is between 1 and 100, as the user must subjectively assess the obtained hypotheses.

Several different quantifiers are available, and they each have their distinct purpose and meaning, for example, *often implies* or *above average correlation*. A comprehensive description of the different quantifiers can be found in (Rauch 2013). The quantifiers can be tuned, so that only the strongest correlations are displayed to the user. If the quantifiers are too weak, the user will obtain an enormous number of hypotheses, most of which will be trivial answers. Usually, the quantifiers need to be adjusted multiple times until a reasonable number of relevant hypotheses are produced.

Once enough relevant hypotheses are obtained, the user can choose one statement at a time for closer inspection. The statement can be evaluated using a contingency table, which shows the strength of the selected hypotheses. This evaluation is purely subjective, as the strength of a hypothesis is case-dependent.

Finally, when a relevant and strong enough hypothesis is obtained, the user can translate its meaning from GUHA- to human language and visualise the results. In this study, this was done manually, but there has been research on automatising the translation and results presentation (Novák et al. 2008).

The statements from the hypotheses are not validated using GUHA and the hypotheses produced by the GUHA method should not be considered as the final answer to an analytical question. A hypothesis is an accepted multivariate correlation within the initial data, but the theory explaining the correlation and causality must be researched with other methods. The purpose of the GUHA method is not to provide these explanations, but to provide novel and interesting correlations that can be researched further.

4 RESULTS FROM THE GUHA METHOD

This study involved querying dozens of different questions, to which hundreds of hypotheses were obtained. However, this paper presents only one of the queries and hypotheses, as the purpose is to demonstrate the capabilities of the method, not the specific outcomes of the analyses. One analytical question was: *What type of track structures are correlated with high TGDRs when the track structure thickness is low and turnout areas are excluded*? This was one of the simpler queries made. Other queries included more complex questions, in which changes in certain parameters were queried and comparative hypotheses were created.

The process of answering the analytical question is presented in Figure 2. First, to translate the analytical question into GUHA language, the antecedents were set so that any track structure features could be chosen. The succedent was set as the observed TGDR. Then, conditions were set to limit the hypothesis only to the parts of the track where the structural thickness was less than or equal to 1.4 m. To clarify, this is considered a low structural thickness in Finland, because the minimum thickness for new track structures is 2.0 m due to frost penetration. Also, turnout areas were excluded from the analysis, as they were found to have vastly different TGDRs compared with line sections. Finally, after a few iterations, the base quantifier was set $a > 500$ to obtain a representable sample and the founded implication p-value was set at 87%.

The data mining resulted in 215 hypotheses. These hypotheses were presented as a list showing their founded implication p-value, antecedents, succedent, and conditions. Using this information, the user could search for a relevant hypothesis and select it for further analysis. When the selected hypothesis for further analysis was translated from GUHA language, it stated:

When structures > 1.4 m in thickness and turnout areas are excluded, the track geometry deterioration rate is high on 89% of structures where the ballast layer MDI is very high, the section is located outside stations, and there is no frost insulation board installed in the track structure.

The antecedents, succedent, and conditions are visualised in Figure 2 on scales with the shaded area depicting their range. The percentage within the shaded range shows the commonness of the range in the whole data set. For example, 98% of the whole data does not contain turnout areas.

The data backing the hypothesis included 512 m (= rows of data) of track along the track section. All structures ≤ 1.4 m in thickness with the turnout areas excluded amount to 10,730 m. At first glance, it might seem the amount of data backing the hypothesis is little by comparison. However, when the subset formed according to the antecedents is examined, it contains some very limiting parameters. The antecedents included high ballast MDI, line sections and the nonexistence of frost insulation boards. Line sections or the lack of frost insulation boards do not limit the original 10,730 m by much, as the selected classes contain over 80% of cases. Conversely, the high ballast layer MDI includes only 6% of the track and greatly limits the subset. If a homogenous area of more than 500 m in total can be constructed with these strict boundaries, the subset can be considered significant.

In this subset, high TGDR values have been observed on 89% of the data. This is exceptionally more common than on the structures ≤ 1.4 m in thickness with turnout areas excluded or on the whole data set, 26% and 20%, respectively. Therefore, it is an abnormality in the data.

This is as far as the analysis can be taken using only the data mining results. For hereon, the analysis requires domain specific

knowledge to put the results into context. In this case, the high ballast layer MDI is particularly interesting as it is uncommon within the data. Furthermore, the common factors causing errors in GPR measurement derived MDI are turnouts and frost insulation boards. As these were both excluded from the analysis, it can be deduced that the high MDIs truly represent high ballast moisture and foulness. Therefore, the hypothesis supports the well-known fact that a fouled or highly moist ballast layer results in deteriorated track geometry. This is an intuitive result that some may even consider self-evident. However, this result shows that the data and the method were suitable for producing a result that has been observed in practice. Therefore, the GUHA method is able to describe real life phenomena from railway data. The commonness of the information obtained from the hypothesis stems from the general nature of the analytical question. If a more specific question is asked, more specific answers can be provided. For example, a more particular query included asking: *How does the TGDR of a structure without a frost insulation board differ from a structure with a frost insulation board when track structures are 1.6 to 2.4 m in thickness (i.e., sufficiently thick regarding frost insulation)*. This question setup was the result of an observation that sufficiently thick track structures should not require frost insulation boards, and yet these were installed in some sufficiently thick structures. Asking these types of questions requires using multiple different quantifiers and GUHA modules. More examples of these types of particular queries, which yielded novel hypotheses, can be found in references Sauni et al. (2020a) and (2020b).

5 DISCUSSION ON THE APPLICATION OF THE GUHA METHOD ON RAILWAY DATA

The GUHA method was found suitable for discovering actual phenomena concerning the load bearing capacity of track structures. The GUHA method works best in cases where some observed phenomenon has multiple different influencing factors, but the relationships among those factors are not known. Unveiling hidden and unusual relationships that are not feasible to investigate by human effort are the GUHA method's main use-case. This makes the GUHA method an applicable method for exploring track load bearing capacity, as there are multiple track structure features jointly influencing it. The GUHA method can highlight which features and their combinations are correlated to observed problem areas. These correlations would be practically impossible to detect from the initial data by human effort. These advantages of the GUHA method can be utilised in practice when a remediation is planned for an old track section and designers have to decide which structures to renew completely and which only partly. In this situation, the GUHA method can be utilised to find the structure types in need for a more comprehensive improvement. A practical use-case for GUHA would be to test multivariate correlations within a data set before using the data in modelling, for example, before using track structure data in modelling the track load bearing capacity.

The limitations of GUHA include the dependency to initial data. Sufficient initial data is required as the initial data is key in obtaining interesting results. The GUHA method does not make projections or models, only descriptions. Hence, a feature that is not depicted in the initial data cannot be included in the results. Furthermore, if the questions regarding the data are too general or the number of influencing factors is low, non-novel answers may be resulted. In these cases, where the relationships between parameters are not complex, typical data pre-processing and correlation analyses provide sufficient methods for correlation analysis. Users of GUHA must keep in mind that the results are only correlations within a given data set, and the results reflect the input data. The validation of the hypotheses obtained from data mining must be done using other research methods. The purpose of GUHA is to produce novel hypotheses to research further using other means.

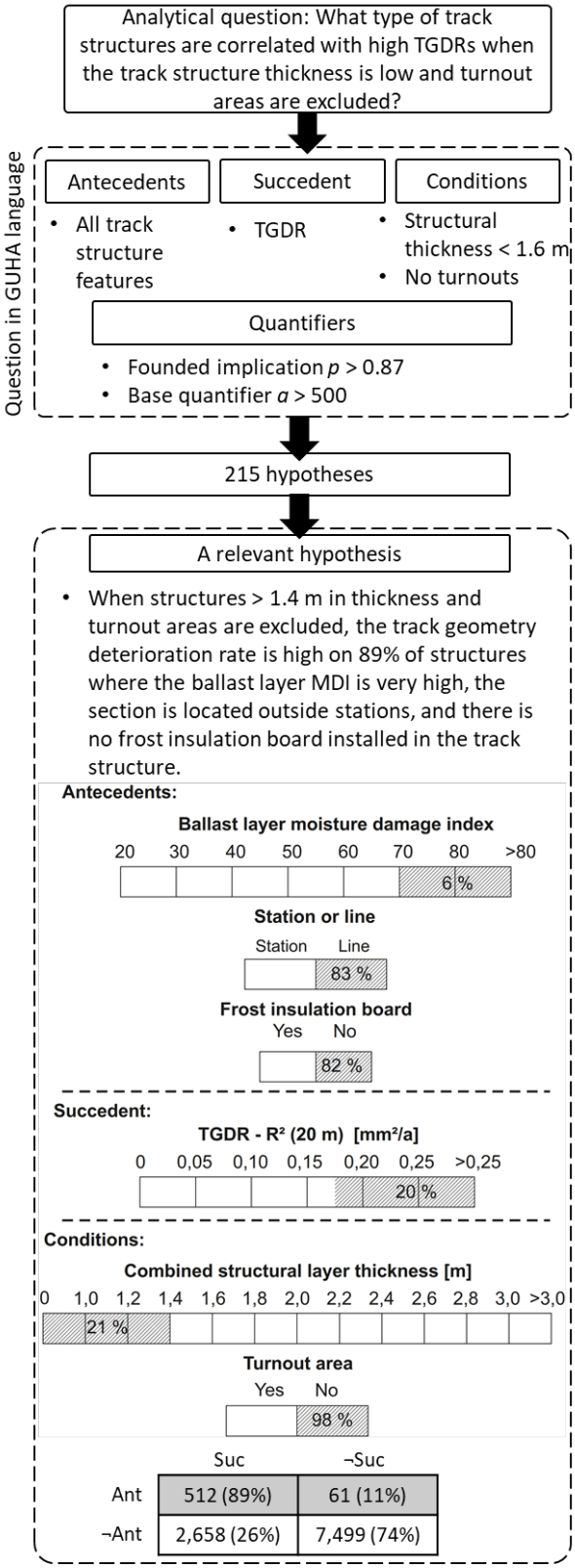

Figure 2. An exemplary GUHA data mining task using LISp-Miner.

6 CONCLUSIONS

This study investigated the applicability of the GUHA data mining method in exploring the causes for realised track load bearing capacity problems using multivariate railway track structure data. The GUHA data mining method describes data in a novel and informative way to the user revealing hidden correlations in the data. The initial data was comprised of several different data sources and depicted different features of the track structure. GUHA data mining was used to search the initial data for correlations between track structure features and realised track load bearing capacity, which was depicted by the TGDR.

This paper reported one exemplary data mining tasks, which demonstrated that the GUHA method could be used to reveal true phenomena from railway data. In practice, the greatest benefit of using the GUHA method is obtained when deciding which structures require more extensive remediation in the next renewal of an old track section. In this design phase, the GUHA method can provide information about the track structure types and features correlated with insufficient track load bearing capacity. This enables frugal use of resources in track renewals.

This study focused on utilizing the GUHA method in investigating a railway track structure data set, but the author believes that there are also other suitable applications for the method from the field of geotechnics. In future research, more cases of linking measured structural resilience data to structure features with complex causality. The GUHA method requires data on an effect and on its possible, obscure causes. Therefore, the GUHA is best used in applied geotechnics. Future research will focus on implementing more initial data sources and quantifiers to investigate the full potential of the GUHA method.

7 ACKNOWLEDGEMENTS

The author wants to thank the Tampere University Foundation Sr and the Finnish Transport Infrastructure Agency (Väylävirasto) for the valued co-operation.

8 REFERENCES

Berka, P. 2016. Practical aspects of data mining using LISp-miner. *Comput. Inform.* 35, 528–554.

Hájek P. and Havránek T. 1978. *Mechanizing Hypothesis Formation: Mathematical Foundations for a General Theory*. 1st Ed. Berlin: Springer.

Higgins, C. and Liu, X. 2018. Modeling of track geometry degradation and decisions on safety and maintenance: A literature review and possible future research directions. *Proc. Inst. Mech. Eng., Part F: J. Rail Rapid Transit*. 232(5), 1385–1397.

Li D., Hyslip J., Sussmann T., and Chrismer S. 2016. *Railway Geotechnics*. CRC Press, Taylor & Francis Group, Boca Raton.

Luomala H., Rantala T., Kolisoja P., and Mäkelä E. 2017. Assessment of track quality using continuous track stiffness measurements. *Georail 2017, 3rd International Symposium Railway Geotechnical Engineering*, Marne La Vallée: IFSTTAR, 281–290.

Novák N., Perfilieva I., Dvořák A., Chen G., Wei Q. and Yan P. 2008. Mining pure linguistic associations from numerical data. *Int. J. Approximate Reasoning*. 48, 4–22.

Rauch J. 2013. *Observational calculi and association rules. Studies in Computational Intelligence*, 1st Ed, Vol. 469, Springer, Berlin.

Sauni M., Luomala H., Kolisoja P., and Turunen E. 2020a. *Determining sampling points using railway track structure data analysis. Proceedings of the 3rd International Conference (ICITG).* Guimarães, Springer, 841–856.

Sauni M., Luomala H., Kolisoja P., and Turunen E. 2020b. Investigating Root Causes of Railway Track Geometry Deterioration – A Data Mining Approach. *Frontiers in Built Environment* 6(122).

Soleimanmeigouni, I., Ahmadi, A. and Kumar, U. 2018. Track geometry degradation and maintenance modelling: A review. *Proc. Inst. Mech. Eng., Part F: J. Rail Rapid Transit*. 232(1), 73–102.