# An Analysis of Track Geometry Data in Combination with Supporting Exogenous Sources Using Linear Regression Techniques

Preece, Joseph; Dean, Ian; Easton, John

*DOI:*
10.1016/j.trpro.2023.11.578



*Document Version*
Publisher's PDF, also known as Version of record

*Citation for published version (Harvard):*
Preece, J, Dean, I & Easton, J 2023, 'An Analysis of Track Geometry Data in Combination with Supporting Exogenous Sources Using Linear Regression Techniques', *Transportation Research Procedia*, vol. 72, pp. 1201-1207. https://doi.org/10.1016/j.trpro.2023.11.578

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 72 (2023) 1201–1207

Transportation Research Procedia

www.elsevier.com/locate/procedia

Transport Research Arena (TRA) Conference

# An Analysis of Track Geometry Data in Combination with Supporting Exogenous Sources Using Linear Regression Techniques

J. D. Preece[a,*], I. Dean[b], J. M. Easton[a]

[a]*University of Birmingham, Birmingham, B29 5JF, United Kingdom*
[b]*Network Rail, Milton Keynes, MK9 1EN, United Kingdom*

**Abstract**

We have investigated relationships between track geometry and weather, to determine if weather has any effect on degradation of track. The data provides an appropriate testbed, covering two Engineers' Line Reference (ELR) IDs of known geological differences, allowing us to explore how weather affects different areas of the railway. We have justified the decision to exploit linear regression modelling, in order to provide a preliminary analysis as the basis for future work. As such, we process and develop a feature table from raw track geometry data that details the track geometry in 200m sections, named Location IDs. From this data, we have applied single and multivariate linear regression models to the dataset and provided an array of visualisations and supporting data. We confirm that linear regression was a suitable investigatory technique, supplying $R^2$ values of up to 69.7%.



*Keywords:* linear regression; track geometry; weather; predictive maintenance

## 1. Introduction

This research forms part of the remit for IN2TRACK2, a project of SHIFT2RAIL, a European Commissioned research portfolio as part of the Horizon 2020 program. The subtask set was to 'Find new signatures in existent data sets that then can improve the performance of the railway system by using artificial neural networks, machine learning.' Network Rail, the UK Infrastructure company of the UK rail network, took lead and commissioned the University of Birmingham, via the UKRRIN network, to provide the technical capability for a joint investigation into railway infrastructure data. Network Rail has over 20 years of digital data from train mounted conditioning monitoring equipment, digital records of asset registry for over 20,000 miles of track, records on asset management interventions, as well as access to their own weather stations across the UK.

In collaboration with the University of Birmingham, this research undertook analysis of the different data analysis techniques to date with available data sets to determine an optimum insight that would support the company's empirical learning of how different routes behave inconsistently to one another and how asset degradation could be


10.1016/j.trpro.2023.11.578

potentially predicted. Network Rail sought to steer the research towards an aspired technical demonstrator where the front-line engineer would be informed of such predictions to then optimise their asset management response and plan effective interventions minimising service affecting failures. IN2TRACK2's phase of this research was to establish the data foundations and understand the 'Art of the possible' concerning scalability vs insights. An essential goal was to achieve results utilising existing data and standard computing hardware as opposed to requiring computer resources not easily available to the front-line such as a supercomputer or cloud computing.

### *1.1. Background*

Climate change detrimentally affects rail infrastructure, with extreme weather events causing failure of rail assets such as bridges, tunnels, turnouts and crossings (Stewart et al., 2020). In particular high temperatures can cause track buckling and bridge deformations; intense and extended periods of rainfall can lead to flooding and landslides that damage embankments, and cause drainage and earthworks failures; and high winds can damage earthworks and other structures (Oslakovic et al., 2012). Outside of extreme events, extended periods of wet or dry weather and variations in temperature can affect ballast settlement and migration (Yeo, 2017). For example, when water in ballast freezes and expands, this can cause movement. The influence of weather on track geometry degradation is significant with studies generally grouped into two categories; temperature and weather events.

Environmental factors are rarely considered in isolation and the majority of these studies incorporate numerous additional factors, notably loading and track design. Elevated rail temperature may cause buckling; a risk that is increasingly likely due to climate change (Dobney et al., 2009). Monte Carlo simulation (Sanchis et al., 2020; Wang et al., 2017), sensitivity analysis (Sanchis et al., 2020; Chapman et al., 2008) and Bayesian network models (Wang et al., 2017) have been applied to determine the impact of temperature on track degradation, with temperature data typically coming from local weather reports, although direct, local, measurement of track temperature is desirable. Another approach is to consider weather events, such as storms, landslips, flooding, falling rock and snow. Data on these events may be obtained from incidence reports or from local weather services.

Algorithms used here include ANNs (Guler, 2014) and Bayesian network models (Wang et al., 2017). Sections of track were segmented and binary fields indicating whether an event e.g. landslip had occurred on a section of track were used as inputs to the algorithm. These data are often used in conjunction with gross tonnage records and track design information. The effect of weather conditions on railway switch ballast degradation and failures has been studied, with average air temperature and level of precipitation forming part of the data studied; linear regression and SVR algorithms were used (Asadzadeh and Galeazzi, 2020).

## 2. Methodology

### *2.1. Scope of Investigations*

This project focused on trying to find a relationship between two datasets; that of the track geometry measurements, and weather. With preliminary analysis of track geometry, we see that there are 109,816,372 lines of track geometry data. Figure 1 illustrates the distribution of the ELR IDs and Track IDs for track geometry. This approximately equal distribution of data is highly suited to our investigations, as it allows us to investigate the difference of the impact of environmental factors against two geologically disparate sections of track. (Though this knowledge is empirical, we envisage that future projects will relate tangible geological data to improve the models.) The LEC1 and LTN1 have total lengths of 133.84km and 183.22km respectively. With this in mind, we can plan our investigations accordingly; investigating how variance in weather affect the degradation of track, and attempting to discover the best level of granularity for the model.

### *2.2. Model Selection*

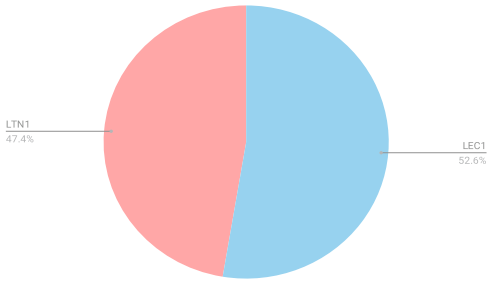
Linear regression provides a solid foundation to build data projects upon, and delivers concrete results which stand alone in their usefulness. With this in mind, we decided to proceed with multivariate linear regression to determine relationships between the data provided. Not only will the results provide useful insights in their own right, but will

allow us to specify the next steps to work towards improved models using advanced machine learning techniques, as part of the next stage in IN2TRACK3.

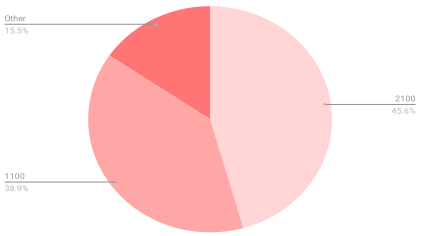
## 3. Results

### *3.1. Preliminary Investigations*

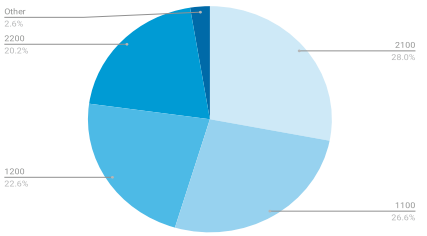
The dataframe needed to create these models can be prepared in approximately 12 minutes and 30 seconds on a commercially available laptop computer, making it feasible for engineering staff without dedicated hardware. From the data provided in summary, we performed a series of linear regression models for various weather parameters (described in Table 1) against the rate of change of the standard deviation. We selected these parameters as the most significant from weather, whilst we selected the rate of change of the standard deviation as the most appropriate measure of track degradation. As such, performing the regression model against these parameters allows us to see how the weather affects the degradation of the track. Moreover, we performed the model over various levels of granularity (described in Table 2) to determine how these levels affected the accuracy of the model.




(a) Distribution between London Euston to Rugby Trent Valley Junction (LEC1) and Liverpool Street to Trowse Lower Junction (LTN1).




(b) Distribution of Track IDs along the LTN1.




(c) Distribution of Track IDs along the LEC1.

Fig. 1. Distribution of ELR IDs and Track IDs for the track geometry data.

Table 1: The four weather parameters selected for investigation.

| Parameter | Unit | Description |
|---|---|---|
| Temperature | °C | The temperature. |
| Relative Humidity | % | The relative humidity, where 100% is total saturation of the air with water vapour. |
| Total Precipitation | mm | The precipitation level. |
| SMI Level | - | The Soil Moisture Index, where 0 indicates a permanent wilting point and 1 indicates that the soil has reached field capacity (wet soil). Values above 1 indicate increasingly wet soils, and values below 0 occur in drought conditions. |

We determine that the most appropriate levels of granularity are that of kilometres and Location IDs, due to the distinctive trend lines and certainties. This section builds on these investigations to compute concrete values from the model and compares them. Table 3 details the $R^2$ values of a sample of sections of track, indicative of the certainty of the model. We see that the finer levels of granularity provide a greater level of certainty, indicating that future models should focus on delivering multiple models for different areas, as opposed to a single generalised model. Furthermore, Figure 2 illustrates the entire spread of $R^2$ values for every kilometre section, and every 200m section.

Table 2: The levels of granularity selected for investigation.

| Granularity | Description |
|---|---|
| All | The entire dataset. |
| ELR IDs | Slices summary into the two ELR IDs. |
| Track IDs | Slices summary into the two ELR IDs, and then the Track IDs. |
| Kilometres | Slices summary into the two ELR IDs, the Track IDs, and then every kilometre of track. |
| Location IDs | Slices summary into the two ELR IDs, the Track IDs, and then every Location ID |

Table 3: The $R^2$ values of a selection of linear regression models, identifying the variance of the model. A higher $R^2$ indicates that the data is well fitted to the trend line.

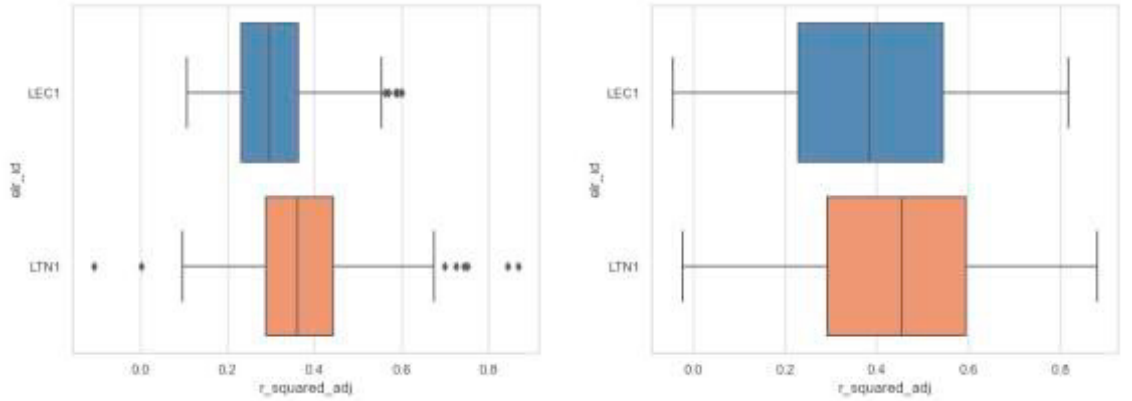
| | ELR ID | Track ID | Kilometre | Location ID | $R2$ |
|---|---|---|---|---|---|
| All | - | - | - | - | 0.195 |
| ELR ID | LEC1 | - | - | - | 0.187 |
| | LTN1 | - | - | - | 0.215 |
| Track IDs | LEC1 | 1100 | - | - | 0.190 |
| | | 2100 | - | - | 0.223 |
| | LTN1 | 1100 | - | - | 0.220 |
| | | 2100 | - | - | 0.242 |
| Kilometres | LEC1 | 2100 | 20 | - | 0.334 |
| | | | 21 | - | 0.299 |
| | | | 22 | - | 0.326 |
| | LTN1 | 2100 | 20 | - | 0.471 |
| | | | 21 | - | 0.558 |
| | | | 22 | - | 0.478 |
| Location ID | LEC1 | 2100 | 20 | 100 | 0.455 |
| | | | | 101 | 0.423 |
| | | | | 102 | 0.336 |
| | LTN1 | 2100 | 20 | 100 | 0.607 |
| | | | | 101 | 0.697 |
| | | | | 102 | 0.470 |

Moreover, Figure 3 illustrates the spread of the trend coefficients for each kilometre section and each 200m section for each of the weather parameters. These results are indicative of the differences between the LEC1 and the LTN1, demonstrating for each weather parameter a shift in effect. For example, Figure 3a indicates that the temperature is likely to have a positive trend for sections on the LTN1, but more likely a negative trend on the LEC1. We attribute these shifts down to the geological disparities between the two ELRs.

## 4. Discussion

This research highlighted the essential task of data foundations and data preparation. To satisfy the requirement of establishing 'new insights', matching asset type, weather condition and an asset condition key metric of degradation provided a sound foundation to build upon for future phases of research. Bringing the different types of weather metrics together with track geometry degradation has proved to yield a relationship worth pursuing. Utilising soil moisture index further established a link with how the track asset behaviour is influenced by how it is supported and serviced. Whilst there is a direct association between soil moisture index with rainfall and temperature; there are other factors such as earthworks construction, drainage, and vegetation as potential further insights into the causality of such relationships. Other additional data metrics yet to be utilised such as line speed, tonnage etc. contribute towards further understanding the causality of track geometry degradation due to the forces and fatigue induced into the track asset.

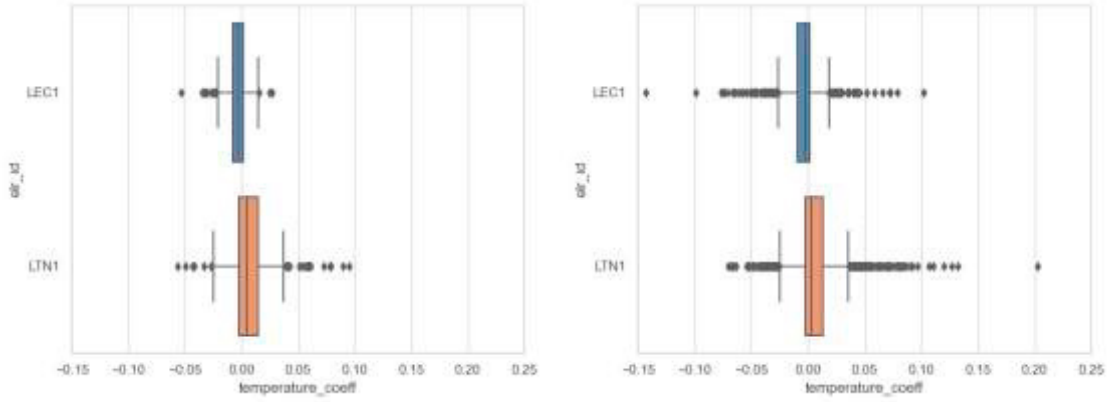
How data is structured for ingestion into the analytics played a critical part in the processing performance and the quality of the insights derived. Not fully utilising the available sampling of track geometry data every 200mm was a key finding and the research corroborated an industry practice of standard deviations for each 200m section as optimum sampling for track geometry asset condition. Preparation works of the remaining data streams being utilised are then the next labour intensive task. It is clear that any form of automation of data analysis using Artificial Intelligence and Machine Learning techniques must be complemented with equal automation of data preparation and error checking.



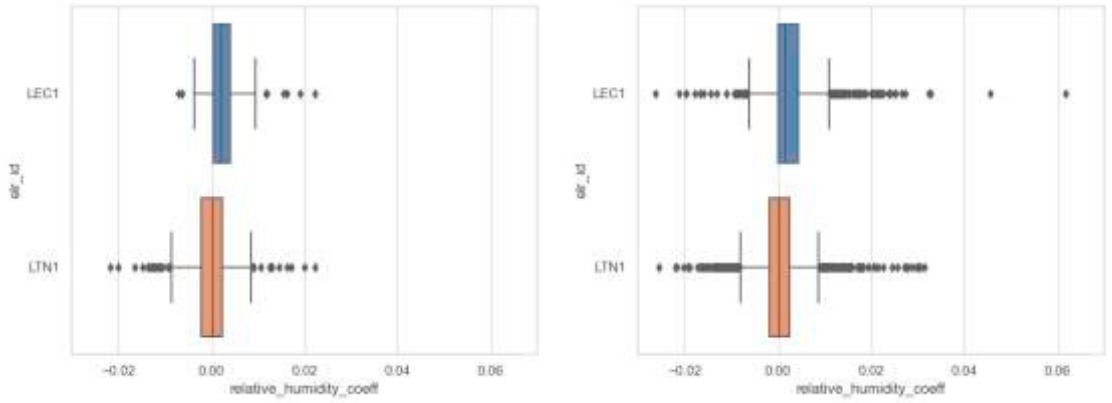

(a) Results of all $R^2$ values for each kilometre along the 2100 and 1100. (b) Results of all $R^2$ values for every 200m section along the 2100 and 1100.

Fig. 2. Results of all $R^2$ values for every 200m section along the 2100 and 1100.

The regression models performed poorly on the larger sections (likely due to the significant variance across such a vast area). It is not until we trained the models on the 1km and 200m sections that the R values began to fall into an acceptable range of values (with the best performing example studied achieving a 69.7% level of certainty without significant outlier removal). This result is an important one; although it shows that we can exploit high-level models with a sufficiently complex model, it is more feasible to provide a lightweight set of calculations from highly localised models. These values will enable track-workers to interpret the data with ease and speed.




(a) The spread of temperature trend coefficients for the LTN1 and the LEC1, for kilometre sections (left) and 200m sections (right).




(b) The spread of relative humidity trend coefficients for the LTN1 and the LEC1, for kilometre sections (left) and 200m sections (right).

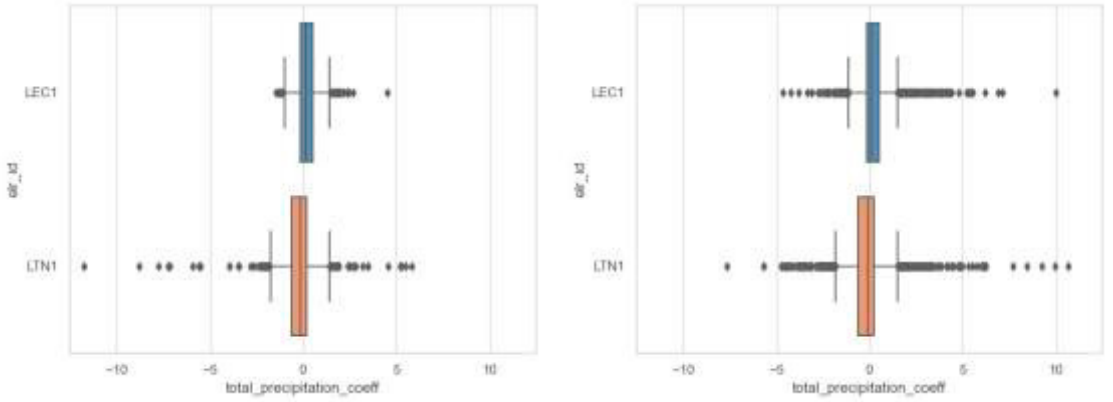
Fig. 3. The spread of trend coefficients for various weather parameters. (1/2)

## 5. Conclusion

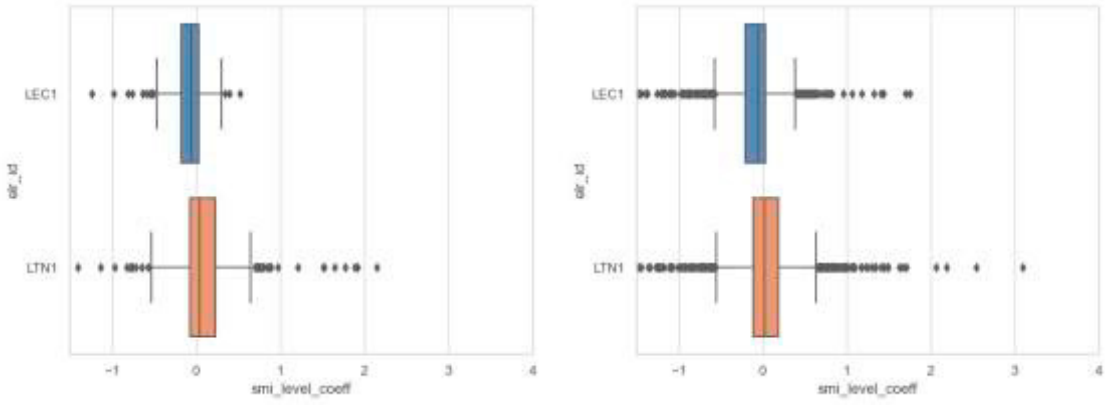
In this paper, we have established a methodological approach that allows the storage, access, and slicing of the core track geometry dataset at arbitrary levels of abstraction, along with available information from associated parameters. We have also demonstrated the applicability of the linear regression approach to the data. The early results suggest that many localised models generated over 1km or 200m sections of track will give predictions of track geometry degradation up to 69.7% certainty.

It is clear that this iterative approach brings scalable, front line derived, insights that are comparable with empirical knowledge concerning differing asset degradation behaviours. With further iterations, increased use of causality data, automatic data cleaning and preparation, and a wider exploration of machine learning techniques predicting degradation should improve, including more complex infrastructure interactions such as ‘Shrink Swell’.

The next steps in the research is to integrate further data streams into the analytics, and to apply more rigorous machine learning techniques to enable predictive modelling for track geometry degradation. This is currently under investigation in the IN2TRACK3 project, bringing in maintenance reports, Integrated Network Model (INM) reports to describe track, and earthworks data. This will validate if the reasoned further causality data, contributes to improved predictions of asset degradation. The balance between data volume ingestion, machine learning techniques and quality of insight should remain scalable for a front-line operator to o plan effective interventions and offset service affecting failures.




(c) The spread of total precipitation trend coefficients for the LTN1 and the LEC1, for kilometre sections (left) and 200m sections (right).




(d) The spread of SMI level trend coefficients for the LTN1 and the LEC1, for kilometre sections (left) and 200m sections (right).

Fig. 4. The spread of trend coefficients for various weather parameters. (2/2)

## References

Asadzadeh, S.M., Galeazzi, R., 2020. An integrated methodology for the prognosis of ballast degradation in railway turnouts. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit 234, 908–924. Publisher: SAGE Publications Sage UK: London, England.

Chapman, L., Thornes, J., Huang, Y., Cai, X., Sanderson, V., White, S., 2008. Modelling of rail surface temperatures: a preliminary study. Theoretical and Applied Climatology 92, 121–131. Publisher: Springer.

Dobney, K., Baker, C., Quinn, A., Chapman, L., 2009. Quantifying the effects of high summer temperatures due to climate change on buckling and rail related delays in south-east United Kingdom. Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling 16, 245–251. Publisher: Wiley Online Library.

Guler, H., 2014. Prediction of railway track geometry deterioration using artificial neural networks: a case study for Turkish state railways. Structure and Infrastructure Engineering 10, 614–626. Publisher: Taylor & Francis.

Oslakovic, I.S., ter Maat, H., Hartmann, A., Dewulf, G., 2012. Climate change and infrastructure performance: should we worry about? Procedia-Social and Behavioral Sciences 48, 1775–1784. Publisher: Elsevier.

Sanchis, I.V., Franco, R.I., Fernandez, P.M., Zuriaga, P.S., Torres, J.B.F., 2020. Risk of increasing temperature due to climate change on´ high-speed rail network in Spain. Transportation Research Part D: Transport and Environment 82, 102312. Publisher: Elsevier.

Stewart, E., Steele, H., Horridge, R., Gonzalo, A.G., 2020. Big Data Analytics for Track Geometry. Technical Report.

Wang, G., Xu, T., Tang, T., Yuan, T., Wang, H., 2017. A Bayesian network model for prediction of weather-related failures in railway turnout systems. Expert systems with applications 69, 247–256. Publisher: Elsevier.

Yeo, G.J., 2017. Monitoring railway track condition using inertial sensors on an in-service vehicle. PhD Thesis. University of Birmingham