Use case report for the Data Infrastructure for National Infrastructure project (DINI)

Forecasting resilience of railway network under propagating uncertainty

Authors: Giuliano Punzo, Ji-Eun Byun, Qian Fu, Iryna Yevseyeva, Tohid Erfani & Konstantinos Nikolopoulos

Contents

1. Foreword

This report refers to the outcomes of the DAFNI sandpit project “Forecasting resilience of railway network under propagating uncertainty” and highlights the areas of the project that could inform the DINI objectives. This is based on the author’s best understanding of the rationale for the report.

1.1 Background and Context

The project “Forecasting resilience of railway network under propagating uncertainty” was funded through the DAFNI sandpit call and aims to compute risks of weather-related disruptions and asset failures in a railway network.

As such, the project is about the Transport sector and involves the research areas of Safety, Transport Security and Infrastructure/Built Environment.

To properly take into account the inevitable effects of uncertainty in railway services, the project focuses on occurrence of uncertainty from weather conditions and asset failures and propagation of uncertainty through interdependent components in a network. To this end, this project collated relevant datasets, develop a computational model to perform probabilistic analysis on network performance, and is making those outcomes available to the public using DAFNI platform.

The project engaged the following team

Giuliano Punzo (PI), The University of Sheffield,

Ji-Eun Byun (Co-I), University of Glasgow,

Qian Fu (Researcher Co-I), University of Birmingham,

Iryna Yevseyeva (Co-I), De Montfort University,

Tohid Erfani (Co-I), University College London,

Konstantinos Nikolopoulos (Co-I), Durham University.

In addition, Samantha Ivings and Marta Zarantonello joined the team as researchers in Sheffield and Glasgow respectively.

The core team then involved stakeholders from Network Rail (Dr Qianqian Li) and planned the attendance to conferences and meetings, including Rail Industry Association (RIA) Consultants Group on climate resilience and adaptation and DAFNI Scotland Roadshow by

Dr Byun, the UK Rail Research and Innovation Network (UKRRIN) Student conference by Marta Zarantonello

Qian Fu will present a poster on “Forecasting Railway Network Resilience under Propagating Uncertainty” at the 11th International Conference on Railway Operations Modelling and Analysis (RailDresden 2025), Dresden, Germany, 1–4 April 2025.

1.2 Description of Activities

The project undertook the following activities:

1. Identifying and collating available datasets for forecasting weather-related disruptions. These have been identified in data already available through Q. Fu at The University of Birmingham. Part of it is made available through DAFNI, in respect to previous constraints about sharing the data, which relate to the acquisition of the dataset by the University of Birmingham.
2. Producing and making available software through the DAFNI platform that can predict the onset of weather-related disruptions by origin-destination, hence by service. This is achieved together and in addition to the specific asset disruptions.

Based on the above, the project produced resilience maps for the UK railway network.

1.3 Benefits of Data Sharing

Weather related asset failures and travel demand were the two datasets mainly used by this work, together with the dataset describing the static railway network topology

As for the dataset related to the weather failures, most data used for the project is kept private, as it was used to train the models, which are delivered as part of the project. Data sharing agreements in place before the project meant that the only a subset of the dataset could and is shared within the project to ensure reproducibility and the possibility of running the model by other users without imputing own data.

These restrictions on the data sharing limit the reach the project could have but are unavoidable in the current breath of the project, for which brokerage of a better data sharing agreement was not possible.

The data sharing agreement in place had the obvious benefit of allowing the development of the project, meaning that posterior failure probabilities of assets could be modelled, and related to the passenger delivery rate.

In this project, the benefits can be categorised under data availability and standardisation. In fact, these are the two main features that allowed the training of the model, albeit not the full sharing of the data, and opened the model to the ingestion of new data, should it become available. The latter was possible by ensuring that the data input stage matches the format in which National Railway, our main point of reference for data, produces its data products.

These two so-categorised benefits are expected to serve the data analytics community as well as the industrial stakeholders that can make good use of the results of the data analysis. In fact, these 2 benefits, when no other barriers are present (such as commercial, cultural, etc.) help a seamless transition from the production to the processing and the use of the intelligence in the data.

1.4 Barriers for Data Sharing

Data sharing in this case required a process (brokerage) that was outside the scope of the project. This identifies the barriers in both the existence of legacy data sharing agreements and the difficulty of brokerage new ones in short time. The latter could be of course overcome by an appropriately sized project, of which data brokerage becomes a fundamental objective beyond being a means to the scope.

Compatibility at the data output-input interface is key to assume smooth transitions between different data environments. This is easily obtained when data move within the same organisation. In the case of this project, data from a variety of sources were formatted in a data product (output) for the software architecture created (input). When the same software was passed onto a different computational architecture (DAFNI) the interface with the data product had to be re-designed.

The problem of data interface is even more pronounced when problem/challenge owners are distinct from data owners and the producers of software for data analysis. In fact, this introduce a third interface, with expectations that could be mismatched between data owners and problem owners, and software producers and problem owners. Possible consequences of this are the production of suboptimal data products or software that matches the data but not the challenge's demands, or vice versa.

1.5 Sources of data – table

Data Source	Data Description	Purpose	Technical Details	Data restrictions and Licence	Barrier	Stakeholder
Name of data, URL if available *Data owner*	*Describe the data that the source provides*	*Describe why it is necessary and the benefit of accessing that data*	- *Data format* - *size* - *APIs metadata description* - *Ontologies* - *Use of Persistent Identifiers* - *Use of Standards*	*Give information on any restrictions on data, and data licences assigned or Data Sharing agreements – share if available.*	*Describe the barrier and assign a barrier from the list*	*Assign which stakeholder(s) this barrier affects*
Weather related asset failures. U of Birmingham through previous data sharing agreement	This dataset contains historical records of wind-related incidents on the Anglia Route of the UK's rail network, along with weather conditions at the time and location of each incident.	Helps identify patterns and correlations between wind conditions and asset failures, enabling predictive maintenance ; Provides historical context to improve rail network resilience to adverse weather; Supports infrastructure risk	Structured, Python-specific pickle file; Metadata description: incident attributes (date, time, location and incident details); weather attributes (wind speed, direction, precipitation, etc.); geospatial metadata (latitude, longitude and	Data is licensed under CC BY-NC-ND 4.0 and restricted to academic or research use only.	Data format may not integrate seamlessly with other datasets; lack of clear documentation or metadata may hinder users unfamiliar with the dataset.	Researchers/academics (Limited access could inhibit studies on weather-related risks and solutions, there may be challenges integrating the dataset into broader predictive models due to format or completeness issues); rail infrastructure managers (may struggle

		assessment, ensuring passenger safety and service continuity.	railway route identifiers).			to use restricted data for real-time operational improvements)
Railway Network Structure Freely available from Rail Data Marketplace (https://raildata.org.uk/dataProduct/P-d6c0c7ee-6743-4999-9b9e-d2dd39585bdb/overview)	A geospatial representation of the UK rail network.	The dataset provides essential information about rail infrastructure (Track) and geographical coordinates.	Shapefile format (.shp).	Data is licensed under OGL3. Subscribers can access the data for free for the duration of their subscription.	It is a link – node model of the rail network but does not contain topological information. Accuracy is variable across the network.	Researchers/academics (Limited access could inhibit studies on weather-related risks and solutions, there may be challenges integrating the dataset into broader predictive models due to format or completeness issues); rail infrastructure managers (may struggle to use restricted data for real-time operational improvements)
Origin Destination passenger demand. Rail data marketplace	A dataframe containing passenger demand between rail network origin/destination pairs.	The dataset provides the number of passengers travelling between all origin/destination pairs for which rail	Pickled file format (.pkl).	Data is licensed under OGL3. Subscribers can access the data for free for the duration of	The travel demand is given in terms of passengers, but not in terms of	See above. In addition, this data may be used to assess the impact of adding/removing rail infrastructure on the delivery of passengers.

		services exist.		their subscription.	services. The latter would be preferable	

1.6 Results Data

Methodologies for data collection / generation

The main data used would be based on the collection, rationalisation and cleaning of existing data. from open sources and partners. The long-term storage and management of the data structure will be part of STFC DAFNI and according to their internal standards. Data description is as in Section 1.5.

Managing, storing and curating data

The team will work locally with research data backed up in near realtime on to the university-owned facilities (e.g. Google Shared Drives, Onedrive, etc.). A repository was created for data and software on Github. Long term management of the data will take place as part of the DAFNI integration as per the specifics of the funding call.

Metadata standards and data documentation

The majority of the input data will be already having suitable metadata attached. The output code and simulation data will include standard textbased metadata documenting the format used, the specification of the numerical simulations, and links to suitable open repositories for the code used/needed for replication.

Suitability for sharing

The data, either used as input or generated as a result of simulations, does not contain any personal information. The input data used is already publicly accessible as open research data or is with the University of Birmingham (contributing through the Researcher Co-investigor). The data will be pre-processed and only the model-ready-to-use data will be shared with the partners and with the research community. The generated simulations will be risk assessed at the point of generation and are planned to be made available for sharing through DAFNI.

1.7 Lessons Learnt and Recommendations

The project identified as good data sharing practice having development environment in which data quality can be set loose for the purpose to speed up software development and proper storage and distribution environment where data are rigorously checked for quality. The project achieved this through sharing the unrefined dataset via different branches in github, for the final version to be shared on DAFNI.

When creating, processing or transferring data as inputs to code, the computational architecture the code uses becomes of primary importance. A delay in the project happened because of the requirement for the software to store data in separate partitions, which could not be accommodated initially on the DAFNI platform.

Contingency resources to work with data are not easy to size. In processing the original data, the project produced intermediate results in the form of datasets the size of which required extra resources for the project.

Data literacy is often disjoint from problem ownership. This is to say that data operators may have only a partial view of the problem they are trying to solve and therefore produce suboptimal data product. Likewise, problem owners often lack the knowledge about the operational environment and envelope in which data are obtained and processed.

An example is the inherent bias that affects the failure dataset of this project. As a dataset of failures in railway assets, there is little to no account of asset availability time, which represent the conditions of the assets most time. This lead to very conservative, i.e. pessimistic, results about the resilience of the railway network.

Addressing this bias means producing a second, wider dataset with additional data points to rebalance the first one. This would be a hybrid dataset, and its creation was not budgeted initially in the project.

The availability of extra funds would be used to produce the expanded dataset to eliminate its inherent bias.