

## A Systematic Review of Artificial Intelligence Public Datasets for Railway Applications

Pappaterra, Mauro José; Flammini, Francesco; Vittorini, Valeria; Bešinović, Nikola

**DOI**

[10.3390/infrastructures6100136](https://doi.org/10.3390/infrastructures6100136)

**Publication date**

2021

**Document Version**

Final published version

**Published in**

Infrastructures

**Citation (APA)**

Pappaterra, M. J., Flammini, F., Vittorini, V., & Bešinović, N. (2021). A Systematic Review of Artificial Intelligence Public Datasets for Railway Applications. *Infrastructures*, 6(10), Article 136. <https://doi.org/10.3390/infrastructures6100136>

**Important note**

To cite this publication, please use the final published version (if applicable). Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.



Review

# A Systematic Review of Artificial Intelligence Public Datasets for Railway Applications

Mauro José Pappaterra <sup>1,2,\*</sup> , Francesco Flammini <sup>2,3</sup> , Valeria Vittorini <sup>4</sup> and Nikola Bešinović <sup>5</sup><sup>1</sup> Department of Computer Science, Uppsala University, 752 36 Uppsala, Sweden<sup>2</sup> Department of Computer Science and Media Technology, Linnaeus University, 351 95 Växjö, Sweden; francesco.flammini@lnu.se<sup>3</sup> School of Innovation, Design, and Engineering, Mälardalen University, 632 20 Eskilstuna, Sweden<sup>4</sup> Department of Electrical Engineering and Information Technology, University of Naples Federico II, 80138 Naples, Italy; valeria.vittorini@unina.it<sup>5</sup> Department of Transport and Planning, Delft University of Technology, 2628 CN Delft, The Netherlands; N.Besinovic@tudelft.nl

\* Correspondence: maurojose.pappaterra.1910@student.uu.se or mauro.pappaterra@gmail.com

**Abstract:** The aim of this paper is to review existing publicly available and open artificial intelligence (AI) oriented datasets in different domains and subdomains of the railway sector. The contribution of this paper is an overview of AI-oriented railway data published under Creative Commons (CC) or any other copyright type that entails public availability and freedom of use. These data are of great value for open research and publications related to the application of AI in the railway sector. This paper includes insights on the public railway data: we distinguish different subdomains, including maintenance and inspection, traffic planning and management, safety and security and type of data including numerical, string, image and other. The datasets reviewed cover the last three decades, from January 1990 to January 2021. The study revealed that the number of open datasets is very small in comparison with the available literature related to AI applications in the railway industry. Another shortcoming is the lack of documentation and metadata on public datasets, including information related to missing data, collection schemes and other limitations. This study also presents quantitative data, such as the number of available open datasets divided by railway application, type of data and year of publication. This review also reveals that there are openly available APIs—maintained by government organizations and train operating companies (TOCs)—that can be of great use for data harvesting and can facilitate the creation of large public datasets. These data are usually well-curated real-time data that can greatly contribute to the accuracy of AI models. Furthermore, we conclude that the extension of AI applications in the railway sector merits a centralized hub for publicly available datasets and open APIs.

**Keywords:** railways; public datasets; intelligent transportation; machine learning; predictive maintenance



**Citation:** Pappaterra, M.J.; Flammini, F.; Vittorini, V.; Bešinović, N. A Systematic Review of Artificial Intelligence Public Datasets for Railway Applications. *Infrastructures* **2021**, *6*, 136. <https://doi.org/10.3390/infrastructures6100136>

Academic Editor: Giuseppe Loprencipe

Received: 30 June 2021

Accepted: 24 August 2021

Published: 22 September 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The automation of traditional manufacturing and industrial practices, by using modern technology in conjunction with massive recollection of data and powerful algorithms, has initiated the course of the fourth industrial revolution. Experts agree that AI is becoming the most central player in Industry 4.0, and the railway industry is not exempt from this. There are many applications within the railway sector in which AI can create a big impact [1]. The increasing number of IoT devices, data available and amount of computer power (along with the decreasing manufacturing costs for technology) create the conditions for the application of modern AI techniques in the railway sector [2].

Applications for AI models are diverse and can be implemented directly in vehicles, infrastructure and services related to transportation [3]. The recollection of data and use

of AI algorithms can aid the analysis of data related to travel routes, the behavior of pedestrians and commuters, the mitigation of energy use, pollution, traffic congestion as well as the improvement of the overall security and safety of passengers. The European Union has assigned over €2.3 billion in funding for the development of smart, green and integrated transport for the period 2014–2020 within the Horizon 2020 Initiative [4]. This initiative comprises different research projects related to the application of AI in transport systems, including the Shift2Rails program, for the development and validation of sustainable, cost-efficient, high-performing, time-driven, digital and competitive train operation standards through railway research and innovation [5].

We recognize several literature reviews on specific AI applications in different railway subdomains. For instance, authors assessed literature on applying Machine Learning to track maintenance [6] and wheel defects [7], and [8] investigated image processing approaches for track inspection. Also, [9] addressed urban flow prediction using machine learning. For traffic management, [10] reviewed data-driven approaches. Similarly, [11] and [12] delved into big data for intelligent transport systems and railway systems, respectively, and [13] looked into potentials of swarm optimization for railways. Differently, [14] reviewed current AI developments across the railway sector holistically, covering all the above mentioned topics as well as safety and security, mobility, and autonomous train driving. They identified that majority of research belongs to maintenance and inspection (57%) and traffic planning and management (25%). The existing literature reviews typically covered a limited scope either regarding specific railway subdomains or some certain aspects of AI, with the exception of [14]. Also, all review papers addressed dominantly AI applications, with little/no focus on the used data. Here, we want to stress that the existence of available data is one of the critical aspects for AI applications. However, authors of AI-focused railway applications rarely publish the related data. In addition, railway companies still tend to be rather conservative and not open to sharing publicly their data to third parties, these being research and academic institutions. Therefore, relevant data are still rather scarce and often one of the largest challenges to address at the beginning of any research effort; according to the findings of the recent survey [15] the lack of suitable datasets for training the ML models has been indicated by the railway stakeholders among the top three obstacles to be faced for the adoption of AI in the rail sector. This situation altogether may lead to the delayed development of AI applications or even prevent researchers from starting to investigate specific topics.

The aim of this paper is to review existing publicly available and open datasets across the whole domain of railway sector for diverse AI applications. We cover subdomains such as maintenance and inspection, traffic planning and management, automated train driving, safety and security, passenger mobility and transport policy. Also, datasets are classified based on its type to numerical, image, label and other. The review covers the period until January 2021. We also present a short overview of supporting public datasets and APIs (Application programming interface (API) is a set of definitions and protocols for building and integrating application software which allows a product or service to communicate with other products and services). All these AI-oriented datasets are of great value for open research and publications related to the application of AI in the railway sector. With this study, we hope to benefit researchers in the fields of computer science and the transport industry by providing an insight into these valuable data and valuable information on how they can be accessed. In addition, companies would hopefully recognize the added benefit and be encouraged to share and publish their data more willingly.

The remainder of the paper is organized as follows. Section 2 describes the methodology utilized for the dataset review. Section 3 presents an overview of selected datasets. Section 4 contains the detailed review of the selected datasets. Section 5 includes a short overview of publicly available supporting datasets and APIs. Section 6 provides the discussion. Section 7 presents challenges and opportunities. Finally, Section 8 presents our conclusions and future research directions. An appendix contains the complete list of reviewed datasets.

## 2. Methodology

### 2.1. Search Criteria and Dataset Selection

For the dataset review, a wide variety of publicly available datasets related to railways were searched. Since no railway-specific databases exist, we used general available and well-known databases, including sources such as: Kaggle, Google Dataset Search, European Data Portal, Data World, Data.gov, Data.gov.uk, Humanitarian Data Exchange (HDX), IEEE DataPort, Zenodo, ScienceDirect's Data in Brief journal and FigShare. These databases are widely used for academic research purposes, some of them implementing quality verifications of the data presented including peer reviews and accompanying publications (e.g., IEEE DataPort, ScienceDirect's Data in Brief). These sources are briefly described in Table 1.

**Table 1.** A short description of all search databases. The databases were accessed on a regular basis between 15 February 2020 and 15 January 2021.

Search Portal	Short Description	URL
Data World	World's largest open data and data collaboration community	<a href="https://data.world/">https://data.world/</a>
Data.gov	Open data from the government of the United States of America	<a href="https://www.data.gov/">https://www.data.gov/</a>
Data.gov.uk	Open data from the government of the United Kingdom	<a href="https://data.gov.uk/">https://data.gov.uk/</a>
European Data Portal	Open data portals across 36 European countries	<a href="https://www.europeandataportal.eu/">https://www.europeandataportal.eu/</a>
FigShare	Open repository of data and papers published in academic research	<a href="https://figshare.com/">https://figshare.com/</a>
Google Dataset Search	Google's dataset search engine	<a href="https://datasetsearch.research.google.com/">https://datasetsearch.research.google.com/</a>
Humanitarian Data Exchange (HDX)	Open platform for sharing humanitarian data maintained by the United Nations (ONU)	<a href="https://data.humdata.org/">https://data.humdata.org/</a>
IEEE DataPort	Dataset storage and search platform maintained by IEEE	<a href="https://iee-dataport.org/">https://iee-dataport.org/</a>
Kaggle	Data science and machine learning portal maintained by Google	<a href="https://www.kaggle.com/">https://www.kaggle.com/</a>
ScienceDirect's Data in Brief journal	Open access journal on published datasets and data articles maintained by Elsevier	<a href="https://www.sciencedirect.com/journal/data-in-brief">https://www.sciencedirect.com/journal/data-in-brief</a>
Zenodo	Open access repository of research papers and datasets maintained by the EU's Horizon 2020 program	<a href="https://zenodo.org/">https://zenodo.org/</a>

The search terms used to find the datasets are single and combinations of the following keywords: "trains", "railway", "rolling stock", "freight", "traffic", "maintenance", "track", "geometry", "signals", "fasteners", "fleet", "management", "safety" and "routing". The inclusion/exclusion criteria for dataset selection were focused on railway data published under Creative Commons (CC), public domain, Open Government Data or any other copyright type that entails public availability and freedom of use. We also excluded datasets that mainly included location, routes or geospatial data of railway assets, or data not specific to railways (such as environmental, geological or weather data). Also, datasets that contained mere statistical data, such as average train speed over a year or average

age of rolling stock, etc., were also discarded. These datasets can be useful in gaining a general insight, and some are shortly mentioned after the review, but cannot be classified as AI-focused. Nonetheless, the focus for the dataset review was strictly concentrated on standalone datasets related to railway applications. In addition, a brief mention of supporting datasets and public APIs is included as a complementary overview in Section 5. We have considered data sources from January 1990 to January 2021. In total, we collected 62 public datasets.

## 2.2. Subdomains in the Railway Sector

This review describes datasets for diverse applications of AI in the railway industry across seven subdomains. The division of railway subdomains is based on RAILS [14]. For each subdomain, we used the related keywords elicited from previous literature reviews on AI applied in the railway sector. The subdomains and the related keywords are as follows.

**Traffic Planning and Management:** This subdomain covers datasets for AI solutions related to planning and scheduling of railway services, rescheduling and fleet management. Keywords include route selection, timetable scheduling, train rerouting and rescheduling, delay management, maintenance scheduling, traffic management, signaling, fleet management, traffic optimization, traffic control, traffic development, infrastructure capacity, capacity allocation, asset management, performance and energy optimization, passenger experience, customer satisfaction and ride comfort.

**Maintenance and Inspection:** This subdomain is reserved to datasets related to the preservation, maintenance and monitoring of railway tracks, assets, infrastructure and rolling stock and communication systems. Keywords include predictive maintenance, condition-based maintenance, corrective maintenance, fault detection and diagnosis (FDD), fault prognosis (FP), wireless communication, train-to-train communication (T2T), and train-to-wayside communication (T2W).

**Safety and Security:** This subdomain includes datasets related to protection against unintended threats (safety) and deliberate threats (security) during railway operations. Keywords include risk assessment, surveillance, cybersecurity, accident prevention, incident response, situational awareness, risk management, system monitoring and abnormality detection.

**Passenger Mobility:** This subdomain is dedicated to datasets that implement AI solutions from the perspective of the mobility of train commuters. Passenger mobility is the ability to move passengers safely and affordably between where they live, work, and spend their leisure time. Keywords include passenger flow, mobility, commuters, and passenger trends.

**Autonomous Train Driving and Train Control:** This subdomain includes datasets that implement AI methodologies to transfer operational responsibilities from manual operators to the train control system and automatic operators. Keywords include automatic train control (ATC), automatic train regulation (ATR), automatic train operation (ATO), automatic train protection (ATP), advanced driver assistance systems (ADAS) and obstacle detection.

**Revenue Management:** This subdomain covers datasets intended to be used in AI applications of disciplined analytics to predict consumer behavior at the different market levels, optimize product availability and design prices to maximize revenue growth. Keywords include prices, tickets, revenue, expenditures and costs.

**Transport Policy:** This subdomain gathers datasets related to dealing with the development of a set of constructs and propositions in order to achieve specific objectives related to social, economic and environmental conditions, as well as to the functioning and performance of the transport system. Keywords include policy, regulations, strategy, environment, land use and equity.

## 2.3. Classification and Types of Data

The datasets are further classified according to four different groups of data type:

**Numerical data**, including track geometry data, train speed data, dynamic data measured from sensors and any other numerical data.

**Image data**, including photos of railway assets, video footage, drone footage and more.

**Label data**, including string data used for classification and cluster-based models.

**Other data**, any other data that falls outside the aforementioned categories, such as simulation instances, 3D scanner data and more.

### 3. Overview of Selected Datasets

For a clear overview, the datasets are divided according to data type, as shown in Table 2. Notice that some datasets might present multiple types of data, dataset (30), for instance, presents both numerical and label data. Table 3 presents the datasets classified according to both railway subdomain and data type. Once again, some datasets might correspond to more than one railway subdomain: dataset (34), for example, falls into two classifications, Traffic Planning and Management and Maintenance and Inspection. Detailed information on each dataset is presented in Appendix A, Table A1.

**Table 2.** Reviewed datasets divided according to data type.

Type of Data	Datasets
Image data	(46,49,50,67,79,81,83)
Numerical data	(16–25,28,30,31,32,34,38,40,42,44,47,51,53,55,57,59,60,65,68,70,72,73,75–77,81,87,89–92,94–100)
Label data	(16–18,20–25,30,31,68,79,81,91,92,94,98)
Other data	(26,27,32–37,60,62,66,85)

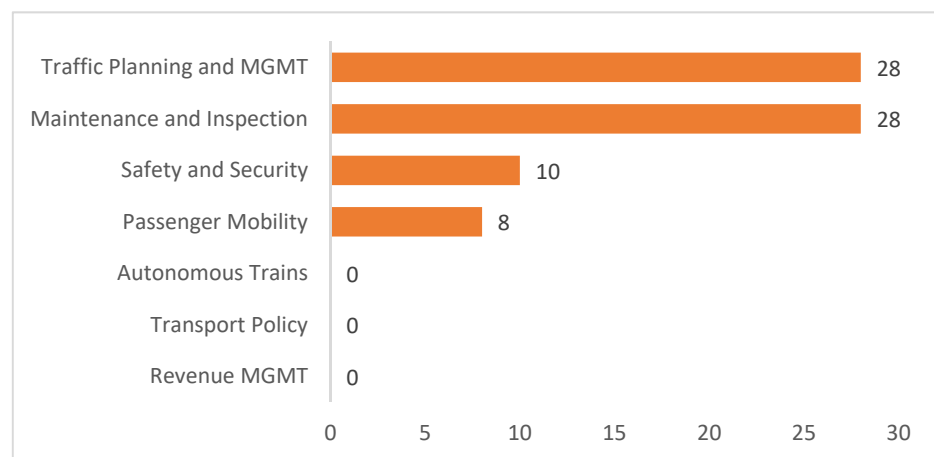
**Table 3.** Reviewed datasets divided according to railway subdomain and data type.

Railway Subdomain	Type of Data	Citations
Traffic Planning and Management	Image data	—
	Numerical data	(16–25,28,30,31,32,38,40,42,95,97–99)
	Label data	(16–18,20–25,30,31,98)
	Other data	(26,27,32–37)
Maintenance and Inspection	Image data	(45,46,49,50)
	Numerical data	(38,34,40,42,44,47,51,53,55,57,59,60,65,68,70,72,73,75–77,87,90)
	Label data	(68)
	Other data	(60,62,66,85)
Safety and Security	Image data	(79,81,83)
	Numerical data	(81,85,89–92,94)
	Label data	(79,81,91,92,94)
	Other data	(85)
Passenger Mobility	Image data	—
	Numerical data	(20,21,95–100)
	Label data	(20,21,98)
	Other data	—

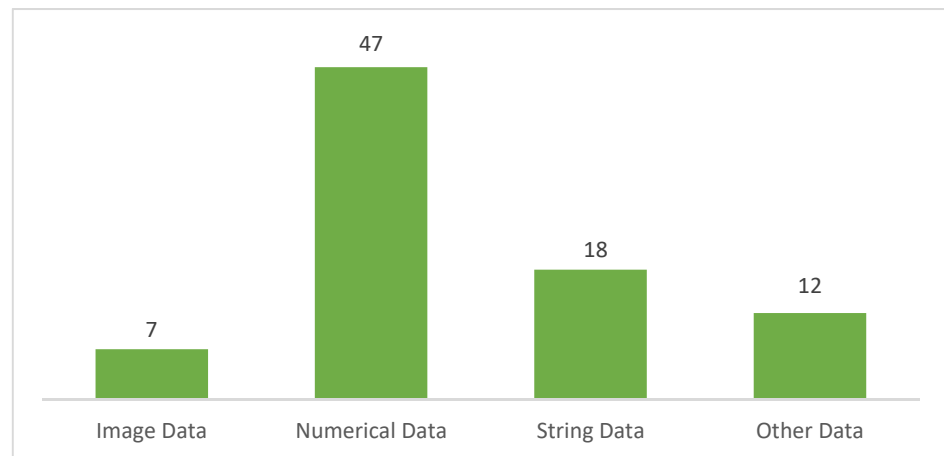
**Table 3.** *Cont.*

Railway Subdomain	Type of Data	Citations
Autonomous Train Driving and Train Control	Image data	—
	Numerical data	—
	Label data	—
	Other data	—
Revenue Management	Image data	—
	Numerical data	—
	Label data	—
	Other data	—
Transport Policy	Image data	—
	Numerical data	—
	Label data	—
	Other data	—

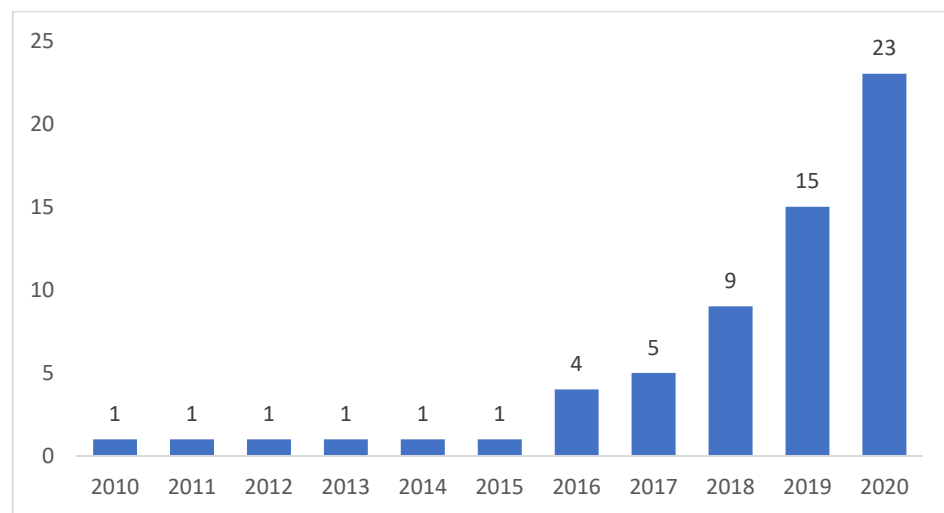
Figures 1–3 present the classification of the reviewed datasets per subdomain, data type and year, respectively. We can observe that the railway applications with the most available datasets are Traffic Planning and Management and Maintenance and Inspection, 28 each; 10 are related to Safety and Security, 8 to Passenger Mobility and none to Autonomous Train Driving and Train Control, Transport Policy and Revenue Management. This shows that Traffic Planning and Management and Maintenance and Inspection are the most prevalent railway domains in AI-oriented studies. Also, numerical data prevails as the most available data type. Finally, there is a gradual growth of openly available datasets from 2016 onwards, with 2020 being the year with the most AI-oriented railway datasets published.



**Figure 1.** Number of datasets (y-axis) divided by railway application (x-axis).



**Figure 2.** Number of datasets (*y*-axis) divided by data type (*x*-axis).



**Figure 3.** Number of datasets (*y*-axis) divided by year (*x*-axis).

#### 4. Review of Selected Datasets

The following sections present reviews of the selected datasets according to the defined railway subdomains.

##### 4.1. Traffic Planning and Management

Within traffic planning and management, we group datasets related to passenger transport, rolling stock and freight transport, and passenger experience.

##### 4.1.1. Passenger Transport

A number of datasets consulted for the review correspond to official government data and datasets that have been facilitated by railway companies. Dataset [16] provides on-time performance (OTP) data on regional trains from the Southeastern Pennsylvania Transportation Authority (SEPTA) of the United States. The data include train ID number, trip direction, origin, date, timestamp and more. GPS data is also provided. Dataset [17] contains data on train delays for the Italian railway line connecting the cities of Bologna and Milan. These data can be used for performance review, timetable rescheduling and the prediction of train delays. Dataset [18] contains timetables of Transport express régional (TER), rail services run by the regional councils of France, with stops and timetables from French railways. This dataset comes from a certified public service and complies with the General Transit Feed Specification (GTFS) format and is created using data collected by the



Department of Regional Trains and Intermodality (Direction des Trains Régionaux et de l'Intermodalité). This dataset can be used for timetable scheduling purposes. Dataset [19] contains real data for train skip-stopping pattern optimization. These data were collected from the Batong line in Beijing's railway network in China. Some of the data contemplated include number of stations, number of trains, capacity of each train, number of passengers, minimal/maximum headway allowed, headway in original train timetable and travel time between adjacent stations. This dataset can be used for timetable scheduling purposes particularly focusing on stop skipping.

Alternatively, some of the datasets have been generated by third parties, utilizing official APIs that are available to the public. Dataset [20] contains granular trip-level performance data on train trips among the NJ Transit and Amtrak railway networks in the Northeastern United States. This dataset includes stop-level, highly detailed data on more than 287,000 train trips, including those extending from the states of New Jersey and New York and covering monthly updates from March 2018 to April 2019. Missing or invalid data from trips are properly reported. Data were obtained using web scraping techniques on available sources from NJ Transit and Amtrak. These data can be implemented for the prediction of train delays and cancellations, analysis of passenger flow and more. Dataset [21] contains data from the commuter train service in the city of Stockholm, Sweden during 2012. This dataset contains timetables and passenger flow data, and it was elicited from the Samtrafik-Trafiklab API. This dataset can be used for timetable scheduling purposes and passenger flow analysis. Dataset [22] contains trip information for the analysis and visualization of Indian Railways. The data include source and destination stations, departure and arrival times, train names and station codes among other information. Similarly to [22], dataset [23] contains timetable data from Indian Railways—the maintainers state that the data is collected from the official Indian government website and dataset [24] contains timetable data obtained from the Indian Railway Catering and Tourism Corporation (IRCTC). Other data include details of trains, and the names of departure and destination stations. These datasets can be used for timetable scheduling purposes.

Other datasets have been created differently, using data collected from computer simulations or hypothetical cases. Dataset [25] contains data on train deviations from planned schedules for resolving the re-scheduling problem, and dataset [26] contains operational data for resolving the schedule optimization and maintenance task scheduling problem. Both datasets were made public by the Institute for Operations Research and the Management Sciences (INFORMS) Railway Application Section (RAS) as part of annual competitions for the application of AI solutions in the railway sector in 2018 and 2016, respectively. These datasets can be used for timetable scheduling and maintenance scheduling purposes. Dataset [27] contains data on four small case scenarios for the timetable scheduling fairness problem. Dataset [28] contains simulation data for the evaluation of performance and delays of a suburban railway line and was published in delay propagation study [29]. The OpenTrack model simulates train departures according to a timetable and calculates speed and positioning of trains while taking into consideration safety rules, signaling and railway line constraints. The characteristics of this simulation data are comparable to other railway networks. This dataset can be utilized for evaluating timetable robustness and performance.

#### 4.1.2. Rolling Stock and Freight Transport

Some of the datasets correspond to operational data and observations from railway traffic. Dataset [30] contains operational data on the daily trips of Australian freight trains to different locations, loading and unloading vehicles in sidings and delivery to final destinations. This dataset can be implemented for timetable scheduling solutions and route optimization models and also for rescheduling and the prediction of train delays. Dataset [31] contains the weekly average number of freight trains held short per day from the Surface Transportation Board's (STB) rail service metrics. The data take into account train type (intermodal, grain unit, coal unit, automotive unit, etc.) and cause

(crew, locomotive, power or other). Data are collected from railroads' daily snapshots and computed in weekly averages. This dataset is maintained by the United States Department of Agriculture (USDA) and can be implemented for freight traffic planning and fleet and crew management.

A number of datasets reviewed were generated using data collected from computer simulations or hypothetical cases. Some of these hypothetical datasets were made public by INFORMS RAS as part of annual competitions for the application of AI solutions in the railway sector. Dataset [32] contains operational data for resolving the train blocking and shipment path (TBSP) optimization problem; similarly, dataset [33] contains operational data for resolving the hump yard classification problem; dataset [34] contains supporting files for modeling railroad yard capacity; dataset [35] contains supporting files for resolving the multi-track territories dispatching problem; dataset [36] contains supporting files for resolving the block-to-train assignment (BTA) problem; and dataset [37] contains operational data for resolving the locomotive refueling problem. These datasets correspond to the annual competitions held on 2019, 2014, 2013, 2012, 2011 and 2010, respectively. These hypothetical data can be used for fleet management and yard management, train forming and scheduling/dispatching.

#### 4.1.3. Passenger Experience

Several datasets are related to passenger experience and ride comfort. Datasets related to measurements of different byproducts of railways such as sound and vibrations can be used to assess passenger experience. Dataset [38] illustrates the effects of train speed and track geometry on passengers' ride comfort. Train operational speeds modify the vibration of the train wagons and reduce ride comfort for railway passengers. This dataset reflects on the combined effect of speed and track geometry on vibration discomfort in high-speed trains, in support of [39]. Dataset [40] contains data measured for the assessment of the influence of vibrations on the noise levels of railway vehicles. Measurements include interior noise in Hz frequency and dB values of two different fasteners at different train speeds over the same non-ballasted track section. Data on exterior noise and vibration spectra for axle box, train floor, vertical track and rail are also provided. This dataset supports the publication on the influence of fastener stiffness on the train's interior noise [41]. A similar dataset [42] contains noise and vibroacoustic measurements for the prediction of rail and bridge noise levels on concrete viaducts. These data are part of the study on rail and bridge noise prediction using multi-layer fastener models among other methods [43]. These datasets could be implemented for improving passenger experience and comfort as well as fault detection and maintenance purposes.

#### 4.2. Maintenance and Inspection

This section includes datasets on maintenance and inspection, and it clusters datasets related to rolling stock, railway track, ballast, catenary and electrical equipment, communication systems and construction works.

##### 4.2.1. Rolling Stock

Using image data can greatly benefit asset identification for the automatization of maintenance tasks. Dataset [44] contains induced failure test data on rolling elements of a spherical roller bearing. The data include vibration records of a bearing, in normal conditions and under rolling-element-induced defects from a test bench experiment. The data are collected within the Railway Technology Research Group (CITEF) of the Polytechnic University of Madrid in Spain. These data can be used for predictive maintenance and related fault detection models. Dataset [45] contains images of pantograph slide plates from various rolling stock on the Swiss Federal Railways' fleet. These images were taken from a rooftop using two high-resolution cameras, with the special aim of capturing the condition of the catenary system.

#### 4.2.2. Railway Tracks

A good number of datasets are related to the maintenance of railway tracks. Image data can be utilized for predicting faults in railway tracks and the identification of parts. Dataset [46] contains images of fasteners from the Dutch railway network. The data is collected by camera-equipped rolling stock that captured the entirety of the Dutch railway network. The dataset was created by the Dutch infrastructure manager ProRail.

The data from the analysis of sleepers and physical phenomena of rolling stock are also relevant. Dataset [47] contains data for the calculation of remaining fatigue life of concrete sleepers on railway systems based on field conditions. A paper published by the authors further explains the methodology used to elicit the data for the prediction of the fatigue life of the asset [48]. This dataset can be implemented for predictive maintenance models and system monitoring. Dataset [49], also created within ProRail and collected with the same methodology as [46], contains images of insulation joints from the railway network. ProRail has also made available a third dataset [50], containing images of insulation joints along with color masks with the aim of detecting spark erosions in the railway tracks. These datasets can be used to train AI models—such as convolutional neural networks (CNNs)—for the identification and localization of railway assets and the prediction of possible failures.

Dataset [51] contains railway track deflection signals obtained from velocity and acceleration measurements. This dataset contains both modeled and measured data for train passages to classify the range of total and downward deflection from train pass-by records. It includes data for a voided sleeper and a good sleeper, and a model for a good sleeper. The records were obtained using inertial sensors; modeled data was obtained using an equation. This dataset was published in a study on automated processing of track deflection signals using velocity and acceleration measurements [52]. This dataset can be implemented for maintenance purposes, including predictive maintenance, fault detection and system monitoring. Dataset [53] contains data elicited from results of different sensitive analyses performed with 3D models. The scope of the analysis includes vehicle speed, vehicle load, number of auxiliary rails and rail pad stiffness. These data aim to identify asymmetric deformations and damaged track components. This dataset supports a study on the use of 3D models for evaluating the use of auxiliary rails in railway transition zones [54]. These data can be implemented for maintenance purposes, including predictive maintenance and fault detection. Dataset [55] contains data for a study published in the Canadian Geotechnical Journal on the evaluation of track stiffness without wheel load data [56]. The data include deflection data, trackside measurements, geometrical data and speed data. These data can be used for system monitoring and for predictive maintenance purposes to quantify and detect changes in the track support stiffness and to determine track deflections. Nonetheless, the dataset maintainers clarify that assumptions need to be made concerning train loading and track behavior due to the unmeasurable variation in train loads. Dataset [57] is part of the study published on the analysis of railway track behavior [58] and contains data on distributed acoustic sensing (DAS) strain and conventional lineside monitoring from strain gauges and digital image correlation (DIC) deflection, as well as other statistical data from the study. This dataset can be implemented for system monitoring and predictive maintenance purposes. Dataset [59] contains historical detection readings for three types of track defects: surface, cross level and dip, and it was made public by INFORMS RAS as part of a collection of open datasets for the application of AI solutions in the railway sector. This dataset can be used for the analysis of track geometry defects for maintenance purposes, including predictive maintenance, fault detection and system monitoring.

The data collected from railway operations can be utilized for maintenance purposes. Dataset [60] contains dynamic responses, GPS positions and environmental conditions of two light rail vehicles in the city of Pittsburgh in the United States, and it is further analyzed in the publication [61]. The data also include acceleration measurements and

track maintenance schedules. This dataset can be utilized for system monitoring, location of assets, predictive maintenance, and scheduling purposes.

#### 4.2.3. Railway Ballast

Data related to the railway ballast can be useful for the maintenance of railway networks. Dataset [62] contains 3D scanner data of two types of railway ballast, calcite and Kieselkalk, including shape analysis information. More information on this dataset is available in its two related publications, a study on shape analysis and ballast stone [63], and a study in discrete element method (DEM) simulations in railway ballast [64]. Another dataset [65] contains measurement data from uniaxial compression tests and direct shear tests conducted on the same two types of railway ballast. For this experiment, direct shear tests were conducted at three different levels of normal load: 10 kN, 20 kN and 30 kN, with three repetition tests for each ballast type and load level. Similarly, dataset [66] is based on cyclic friction tests of ballast stones interfaces with different vertical loads. The data are the result from a study on cyclic ballast–ballast friction tests [67] under varying loads expressed in coefficient of friction (CoF); and is presented as 3D scanner data of angular-tip stones (before and after the tests were conducted). These datasets can be utilized to assist computer-based simulations and AI models intended to maintain railway ballast, for predictive maintenance purposes and to ensure the safety of railway operations.

#### 4.2.4. Catenary System and Electrical Equipment

Dataset [68] contains data published in a Chinese study on wind-induced responses of the catenary for a high-speed railway [69]. This dataset illustrates the dynamic behavior of the catenary system in different flow conditions (turbulent and uniform) in four tension combinations by measuring the displacement ( $D2 \times$ ) and acceleration ( $A1 \times$ ,  $y$ ) at the hanging point of the steady arm. The data are collected with the implementation of a micro-acceleration sensor with high sensitivity and a laser displacement meter. These datasets can greatly optimize different maintenance tasks including predictive maintenance, fault detection and system monitoring.

Data collected from electrical measurements can be utilized for the maintenance of railway electrical equipment. The dataset of measured and commented pantograph electric arcs in DC (direct current) voltage railways [70] presents digitized sampled data using a data acquisition system located onboard and connected to voltage and current sensors. The data were collected from Trenitalia (Italy) and Metro de Madrid (Spain) in late 2019 for the international research project MyRailS [71]. The data include recordings of pantograph electrical quantities which are difficult to obtain from real-world operations. A similar dataset [72] contains data of measured pantograph voltages and currents of European AC (alternating current) railways. The set contains digitized sampled data using various data recorders located onboard and connected to voltage and current sensors (voltage dividers and Rogowski coils). The data were collected from four major AC European railways including the Zurich–Brig railway line (Switzerland), the Rome–Naples railway line (Italy), the Hamburg–Dortmund–Frankfurt railway line (Germany) and the Paris–Lyon railway line (France). Both these datasets can be used for different maintenance purposes, including power quality studies, interaction of rolling stock with the overhead contact line, analysis of electrical phenomena of the systems, optimization of power traction supply in DC/AC rolling stock as well as the assessment of other energy-related areas (such as signaling). The  $2 \times 25$  kV Railway Feeding System Simulation Database [73] set contains electricity data measurements from computer simulations. The simulation is based on a simplified model of a  $2 \times 25$  kV bi-level traction power system for feeding high-speed trains; this model is defined by a study that presents traction system models and solvers for extensive network simulations [74]. This dataset can be utilized for the assessment of electrical equipment in high-speed railways.

#### 4.2.5. Communication System

Dataset [75] contains performance data of transmission control protocol (TCP) congestion control algorithms in high-speed railway scenarios and in mobile and static scenarios from computer simulations. The algorithms measured are Hd-TCP and TCP NewReno. These data can be utilized for the assessment, monitoring and fault prevention of long-term evolution (LTE) communication networks in railway systems.

#### 4.2.6. Construction Works

Data collected during railway restoration and construction work can be utilized for maintenance purposes. Dataset [76] includes monitoring data for a railway bridge before, during and after a retrofitting process. The bridge is located in KW51 in Leuven, Belgium. These data include different measurements of the structure during the process, such as acceleration on the bridge deck and the arches, strain on the bridge deck and the diagonals connecting the bridge deck with the arches, strain on the rails, displacement at the bearings and also includes measurements on the temperature and relative humidity of the area. Similarly, dataset [77] contains data collected during the rehabilitation of the Buna railway bridge in Croatia. The dataset contains acceleration data corresponding to a roving test, with specified positions and orientations from an experiment conducted during work on the bridge to implement ultra-high-performance fiber-reinforced concrete (UHPFRC). This dataset supports the study published on the rehabilitation of the Buna bridge [78].

### 4.3. Safety and Security

This section presents datasets related to railway safety and security and it combines datasets into situational awareness, surveillance, accident prevention and risk assessment.

#### 4.3.1. Situational Awareness

Image data can be utilized to better understand the railway environment and further promote safety and security. The Cityscapes dataset [79] contains a set of diverse stereo video sequences recorded in public street scenes from 50 different cities in different times and under different weather conditions. This dataset includes high-quality pixel-level annotations of 5.000 frames and a complementary 20.000 weakly annotated frames. Other features include polygonal annotations with dense semantic segmentation for vehicles and people. The level of complexity of this dataset is high, comprising 30 different classes clustered in eight different groups, including a class for railway vehicles. The labeling policy and other rich metadata are available to the public. More information on this dataset is available in the dataset publication paper [80]. The RailSem19 dataset [81] contains a set of 8.500 images taken from a rail vehicle perspective for semantic rail scene understanding. The dataset includes extensive semantic annotations, based on geometric (polygons and polylines) and dense label maps. A good number of image frames show intersection zones of road and rail vehicles as well as other complex railway scenarios. Difficult weather and lighting conditions are also taken into consideration. Some of the labels presented in RailSem19 are compatible with the Cityscapes dataset [79]. More information on this dataset is published in the supporting paper for RailSem19 [82]. Possible applications of these datasets in the railway sector include safety and security, surveillance, situational awareness and any computer-vision-based models that interact with railway surroundings and city environments.

#### 4.3.2. Surveillance

Some of the datasets were constructed with surveillance as their main purpose. The PETS 2017 dataset [83] contains data from on-board surveillance systems intended to protect critical assets. PETS stands for performance evaluation of tracking and surveillance, and its application is intended to evaluate the performance and detection of various surveillance events. This dataset contains two different sets of data: the ARENA dataset and IPATCH dataset. ARENA includes 22 scenarios captured on multicamera RGB video recordings

for the detection and understanding of human behavior around a static vehicle. The focus of these acted scenarios is the classification of normal, abnormal/rare and threatening behavior. The IPATCH dataset presents piracy-inspired scenarios and implements not only image data but data from various sensors, but it is aimed at waterborne vessels. Further details on the dataset can be found in its publication [84]. PETS 2017 can be implemented for surveillance purposes in railway stations and platforms. Dataset [85] contains a 3D point cloud of a railway trench in Lavancia-Épercy in France. The 3D point cloud is composed of 110.356.682 million points containing X, Y, Z and RGB information. This dataset was conceived within the multi-scale observation and monitoring of railway infrastructure threats (MOMIT) initiative backed by the European Union [86]. The dataset is intended to observe and monitor railway infrastructure threats and can be used for both security and monitoring purposes.

#### 4.3.3. Accident Prevention and Risk Assessment

The observation of physical and geological phenomena surrounding railway environments is crucial for maintaining operation safety. Dataset [87] is designed as a monitoring and early warning method for a rockfall along railways, based on the characteristics of vibration signals. The dataset was generated with the results of a rockfall test in which the vibration signals of rocks falling over a flexible safety protection net and different areas of the rails were obtained [88]. This dataset can be implemented for the development of an early warning system for the safety and security of railway network operations. Similarly, dataset [89] contains data from experiments on granular flow behavior and deposit characteristics. The data contained include relationships between mean grain size and global shear as well as other numerical data from physical and geological phenomena. These data can be used to infer implications for rock avalanche kinematics, which is relevant to the railway sector. Another example is dataset [90], which includes geomagnetic and geoelectric field values generated by analytic calculations. In this dataset, two sets of data are provided: the first set contains data sampled once a second and containing seven frequency components, and the second set contains data sampled once a minute and containing six frequency components. Geoelectric fields generated from the variation in geomagnetic files can affect the operation of railway circuits. These datasets can be implemented for risk assessment, safety purposes, situational intelligence and, predictive maintenance.

Historical data and social engineering data can be used for the prediction and early detection of accidents. Dataset [91] contains data on traffic accidents that occurred in France from 2005 to 2016, including railway accidents. This dataset contains data on type of collision, atmospheric conditions, surface condition, people involved, vehicle information and many other factors. Dataset [92] contains data on over 3500 animal–train collisions and over 10.000 locations provided by Polish State Railways (Polskie Koleje Państwowe, PKP). The data ranges from 2012 to 2015. Some of the traffic characteristics and factors considered in the dataset include traffic intensity, speed, rail curvature, land use and animal habitat characteristics. These data support a study on the relationship between animal habitat composition and population and ungulate–train collisions across the country [93]. Dataset [94] contains data on the work schedules and sleep patterns of railroad employees from a study sponsored by the Federal Railroad Administration (FRA) in the United States. The aim of this study is the analysis of work-schedule-related fatigue in railway employees via the documented work/rest schedules and sleep patterns of a test group. This dataset includes work schedule and sleep pattern data of signalmen, maintenance of way (MOW) workers, dispatchers, and train/engine service workers in both freight and passenger trains. These datasets can be implemented for safety and risk assessment purposes.

#### 4.4. Passenger Mobility

This section contains datasets related to passenger mobility, mostly on passenger flow estimations.

### Passenger Flow Estimation and Trends

Datasets that reflect on passenger trends and occupancy levels can be implemented for estimating passenger mobility in order to optimize train operations. Some of the datasets consulted contain data on passenger trends from rail operating companies (ROC). For instance, dataset [95] contains ridership data on the Bay Area Rapid Transit (BART) network in the San Francisco Bay Area railway network in northern California (United States). This dataset includes hourly ridership divided by year, starting from 2011, and was generated using ridership reports provided by the government-owned company. These data can be implemented for the prediction of train occupancy, passenger flow and infrastructure capacity. Dataset [96] contains passenger frequency data from Swiss Federal Railways (SBB-CFF-FFS). The data were collected during operations in 2014. These data can be used for passenger flow estimation. Dataset [97] contains data captured from Deutsche Bahn (DB) trains and travels in different stations in Germany. These data were compiled from online information available from the DB service, and can be used for timetable scheduling, passenger flow analysis and traffic planning purposes. Dataset [98] contains data from metro lines in India for the prediction of traffic and passenger flow. The dataset contains a training and test sets for the purpose of predicting traffic volume according to weather conditions. Some of the features included in this dataset are the date, pollution index, humidity, wind speed and direction, visibility, snow, clouds, weather description and corresponding traffic volume in the metro. These data can be implemented for the prediction of train occupancy, passenger flow and infrastructure capacity.

Other datasets contain data that are not connected to a particular railway company. Dataset [99] contains monthly data on train occupancy from 1999 to 2011. These data can be implemented for the prediction of train occupancy, passenger flow and transport capacity. Dataset [100] present records of crowd density on several trains over a span of months. Some of the values in this dataset have been altered or obscured for security reasons. This dataset can be implemented for the analysis and understanding of patterns in passenger flow.

### 5. Supporting Datasets and Public APIs

A number of general datasets and APIs—that are publicly available—may not be directly applicable for intelligent railway models as standalone data. Nonetheless, they can be very useful for various reasons and can be used as complementary data. This section presents an overview of some supporting datasets and public APIs that were found. Table 4 presents the list of publicly available supporting datasets and public APIs. The number of datasets portrayed were limited to sixteen for illustrative purposes.

**Table 4.** Selected supporting datasets and public APIs. The links were accessed between 15 September 2020 and 15 January 2021.

Name/Title	Description	Last Updated	Link
ERAIL-Railway accident and incident links	Reports of railway accidents and incidents within the European Union	2020-10	<a href="https://data.europa.eu/euodp/en/data/dataset/erail-investigations">https://data.europa.eu/euodp/en/data/dataset/erail-investigations</a>
Train Stations in Europe	Names, coordinates and properties of European railway stations	2020-10	<a href="https://www.kaggle.com/headsortails/train-stations-in-europe">https://www.kaggle.com/headsortails/train-stations-in-europe</a>
Grade Crossings Inventory	An inventory of the location and characteristics of railway crossings in Canada	2020-09	<a href="https://open.canada.ca/data/en/dataset/d0f54727-6c0b-4e5a-aa04-ea1463cf9f4c">https://open.canada.ca/data/en/dataset/d0f54727-6c0b-4e5a-aa04-ea1463cf9f4c</a>
Railroad Crossings	Detailed information of all railway crossings in the United States	2020-05	<a href="https://hifld-geoplatform.opendata.arcgis.com/datasets/railroad-crossings">https://hifld-geoplatform.opendata.arcgis.com/datasets/railroad-crossings</a>

Table 4. Cont.

Name/Title	Description	Last Updated	Link
HARCI-EU	Geospatial data of critical infrastructure, including railway networks	2019-12	<a href="https://figshare.com/articles/dataset/HARmonized_grids_of_Critical_Infrastructures_in_Europe_HARCI-EU_/777301">https://figshare.com/articles/dataset/HARmonized_grids_of_Critical_Infrastructures_in_Europe_HARCI-EU_/777301</a>
Citylines: Transit systems of the world	Transportation line data from cities from around the world, and historical data of railway line development	2019-03	<a href="https://www.citylines.co/">https://www.citylines.co/</a>
National Railway Network-NRWN-GeoBase Series	Geometric descriptions and basic railway attributes	2018-05	<a href="https://open.canada.ca/data/en/dataset/ac26807e-a1e8-49fa-87bf-451175a859b8">https://open.canada.ca/data/en/dataset/ac26807e-a1e8-49fa-87bf-451175a859b8</a>
Freight Analysis Framework (FAF)	Flows of goods among US regions for all modes of transportation, including railways	2017-08	<a href="https://www.kaggle.com/usdot/freight-analysis-framework">https://www.kaggle.com/usdot/freight-analysis-framework</a>
WFP-Global railways	Geodata about global railways	2017-05	<a href="https://data.humdata.org/dataset/global-railways">https://data.humdata.org/dataset/global-railways</a>
Rail Network	Linear network representing railway tracks and other data from Canadian railways	2012-06	<a href="https://open.canada.ca/data/en/dataset/c2c4f386-a736-4eaa-b5b6-28c3a8f75466">https://open.canada.ca/data/en/dataset/c2c4f386-a736-4eaa-b5b6-28c3a8f75466</a>
Railroad Bridges	Detailed information on all railway bridges in the United States	2009-09	<a href="https://hifld-geoplatform.opendata.arcgis.com/datasets/railroad-bridges">https://hifld-geoplatform.opendata.arcgis.com/datasets/railroad-bridges</a>
National Rail Enquires (NRE)	Real-time train data in Great Britain as a public API	Active	<a href="https://www.nationalrail.co.uk/46391.aspx">https://www.nationalrail.co.uk/46391.aspx</a>
Sydney Trains Service Interruptions RSS Feed	Real-time machine-readable feed of Sydney Trains information concerning service interruptions	Active	<a href="https://opendata.transport.nsw.gov.au/dataset/public-transport-realtime-alerts-0">https://opendata.transport.nsw.gov.au/dataset/public-transport-realtime-alerts-0</a>
Nederlandse Spoorwegen (Netherlands Railways)	A public REST API for Dutch Railways in the Netherlands	Active	<a href="https://www.ns.nl/en/travel-information/ns-api">https://www.ns.nl/en/travel-information/ns-api</a>
Traffiklab	A collection of APIs for public transport in Sweden	Active	<a href="https://www.trafiklab.se/">https://www.trafiklab.se/</a>
Google Transit APIs	Google's API services for GTFS Static and GTFS Realtime	Active	<a href="https://developers.google.com/transit">https://developers.google.com/transit</a>

A number of government websites provided information concerning railway assets and configurations. In North America, for example, the Rail Network dataset contains data on Canadian railways, including a linear network that represents railway tracks and other rail data such as geometry, operator's name, owner's name, track condition, subdivision name. These data can be implemented for traffic management, maintenance and other purposes. The National Railway Network (NRWN) GeoBase Series dataset contains geometric descriptions and basic attributes of Canadian railway systems. The data include tracks, junction, crossings, marker posts, stations and structures that are associated with descriptive attributes (e.g., track classification, operator, gauge and others). This dataset can be implemented in different areas including maintenance, security, system monitoring and more. Moreover, the Grade Crossings Inventory dataset contains an inventory of the locations and characteristics of railway crossings in Canada. Some of the data provided include location (latitude and longitude), responsible road authority, previous number of accidents, fatalities, injuries, number of daily trains and vehicles, and other data such as train max speed (mph), road speed (km/h) and more. This dataset is maintained by the Government of Canada, and it can be implemented for risk assessment and safety purposes. The datasets on Railroad Crossings and Railroad Bridges available on



the Homeland Infrastructure Foundation-Level Data (HIFLD) website contain the location of and other detailed data on train assets around the United States. The first dataset contains data on more than 86,000 railroad bridges, including a large number of attributes depicting location, classification, geometrical data, description and more. The second dataset contains detailed information on more than 245,000 railway crossings including location, railway line, city and security reports among other data. The US Department of Transportation created the Freight Analysis Framework (FAF) dataset. This dataset represents flows of goods among regions in the United States for all modes of transportation, including railways. This dataset intends to map the flow of freight transportation across the country by combining data from a variety of sources including the 2012 Commodity Flow Survey (CFS), international trade data from the Census Bureau and data from other industry sectors including agriculture, resource extraction, utility, construction, services and more. This sort of dataset can be implemented for the recognition of transport trends and patterns.

European governments also provided extensive supportive datasets. The HARMONIZED grids of Critical Infrastructures in EUROPE (HARCI-EU) dataset contains data on harmonized grids of critical infrastructures (CIs) within the European Union [101]. CIs are defined as infrastructures essential to the safety and well-being of people. This dataset contains geospatial data of CIs, including railway transport infrastructure. Specifically, the dataset consists of 22 grids in GeoTIFF format with a resolution of 1 km<sup>2</sup>. HARCI-EU uses the ETRS89 coordinate system and a map that implements a Lambert azimuthal equal-area projection scheme. The spatial distribution of railway infrastructures and their economic value can be implemented for risk assessment, safety, and security purposes. The European Railway Accident Information Links (ERAIL) database presents updated information on railway accidents and incident reports across the member states. The data can be filtered by country and includes information such as date, location, number of fatalities and injuries and investigation details of more than 3,000 events. In an effort to map the extension of European railways, the dataset contains the names, coordinates and basic properties of more than 36,000 train stations located in or adjacent to European territory. This dataset includes location data (latitude–longitude), country and city names, UIC code and other data.

On a worldwide scale, the United Nations World Food Programme (WFP) compiled a dataset depicting all railway systems worldwide. This dataset contains more than 110,000 entries and contains location coordinates, geometrical data and characteristics of railways worldwide. The Citylines dataset contains a complete description of railway networks from more than 350 cities from around the world, including the historical data of railway line development. The data presented include geospatial data, systems, lines, sections, stations and more.

In addition, there is a large number of public APIs that provide data on railway operations. These APIs are often maintained by government organizations, transport agencies and railway operators, and can be utilized as complementary data for intelligent models. The British National Rail Enquires (NRE) API provides an open data feed on TOCs across England, Scotland and Wales. It utilizes three engines: Darwin, Knowledgebase (KB) and Online Journey Planner (OJP). Darwin provides timetable information, including departure predictions, timetable rescheduling, service cancellations, predictions of delays and historical data for the previous twelve months. KB provides data on the UK railway network, including static data, such as information on station facilities, and real-time information, such as service disruptions and engineering work. Lastly, the OJP engine provides data on routes, fares and availability for planning purposes. The Sydney Trains Service Interruptions RSS Feed contains a real-time machine-readable feed of train information concerning service interruptions in the trains in Sydney (Australia). The alerts function on the stop, trip or service line level and are provided in a GTFS format. The Dutch TOC Netherlands Railways (Nederlandse Spoorwegen, NS), provides REST API to handle a large amount of data on timetables, prices, live departure times, service disruption and engineering work, and it also includes geodata on all stations in the Netherlands.

Other available APIs have been constructed by combining different publicly available APIs and RSS feeds that are related to public transportation. Traffiklab comprises a collection of APIs for public transport services in Sweden, including regional real-time GTFS data, vehicle positioning of transport vehicles, disturbances and interruption information, traffic information and data on stops and assets from Swedish railway companies. In general, General Transit Feed Specification (GTFS) (General Transit Feed Specification (GTFS), <https://gtfs.org/> (accessed on 15 August 2021)), known also as Google's Transit API, provides tools for transit companies to share static and real-time data, including routes, stops, trips, service alerts and schedules. It utilizes two API extensions, GTFS Static and GTFS Realtime, for static and real-time data, respectively.

## 6. Discussion

As we have previously established, the possible applications of AI in the railway sector are vast, and there are many railway challenges and tasks that can greatly benefit from AI-based models. The heart of most of AI models is the data that is used to feed its logic. Looking at publicly available datasets we recognized some promising sources that can be used for research without any constraints regarding the publication of results.

Most of the publicly available datasets were related to Traffic Planning and Management. The most common types of data found in publicly available datasets are numerical and string data. For planning and rescheduling, additionally, some datasets have been generated by third parties using official APIs as the main source of data. Other datasets have been generated from computer simulations and hypothetical cases. These datasets contain timetable information, passenger flow data, trends and occupancy levels, information on delays and timetable performances. There are also other less common data such as operational measurements on speed, vibration and sound that can be used for the evaluation of passenger comfort onboard the trains.

When it comes to Maintenance and Inspection, there were more publicly available datasets than originally thought. There are open data on different railway maintenance applications. Some of the data were collected on train operations using IoT devices and smart sensors. The data reviewed can be clustered according to the train and railway's physical characteristics, these aspects including railway tracks, sleepers and ballast, electrical and communication components. Other data were collected from the vibration generated during railway restoration and construction work.

There are few openly available datasets related to the area of Safety and Security. Some public image datasets on traffic can be adapted for railway applications related to situational awareness. These datasets are complete and well documented. The same is true in the case of historical accident data. Nonetheless, there were no public datasets specific to railway security that were openly available under CC copyright or similar. Even though they were not specifically designed for railways, some datasets containing measurements on geological observations are important for both the safety and maintenance of railway networks. These data are of great importance for risk assessment and predictive maintenance purposes.

Comparing the existing AI applications reviewed in [14] and the public datasets, on one hand, even though that majority of applications is focused on maintenance and inspection of 57%, this is not reflected in the corresponding datasets—only 25%. On the other hand, regarding traffic planning and management, significantly higher proportion of datasets is available, there we see 25% of applications vs 50% of datasets. Also, [14] determined 8% of papers related to Autonomous train driving, but no public datasets were found; and similar holds for Passenger Mobility. Finally, following the results obtained in [14], no public datasets were found related to Revenue Management and Transport Policy.

Supporting railway datasets and APIs that are openly available to the public are vast and easily found. For the purpose of this review, we have only mentioned some illustrative examples of these data. Nonetheless, these supporting sources could be used to

aid AI-based models by reading real-time data related to many railway application sectors. These data can be implemented in AI models for complementary purposes. Furthermore, real-time data can be harvested from these publicly available sources in order to create new datasets that can later be used to feed AI models. Such complimentary datasets would provide more system-specific characteristics and lead to more accurate AI-based solutions.

## 7. Challenges and Opportunities

This section discusses the shortcomings and the possibilities of the reviewed datasets, along with some observations from the authors' perspective.

First, the main challenge for this review was the low number of publicly available datasets and its uneven distribution over subdomains. Considering the amount of literature regarding the application of AI in the railway industry as reviewed in e.g., [6,10,12,14], the number of publicly available datasets is very narrow in comparison. Additionally, it was not possible to obtain any dataset implemented in the previous literature reviews [14]. Unfortunately, most datasets are still private or not copyrighted as CC or open data, which render their utilization in public research limited or even impossible. Second, the lack of suitable image datasets is evident. Out of 62 datasets reviewed, only six image datasets could be found that were related to the railway industry and openly available to the public. This is a big gap considering the amount of research done involving computer vision models in the railway sector. As seen in the previous literature reviews, the collection of these data takes time and effort, which could explain why these datasets are not made publicly available by researchers. Data privacy issues could possibly be related to image datasets. Third, another big challenge encountered was the lack of proper documentation and metadata information on publicly available datasets. Only a very small number of the datasets contained detailed documentation, and only a small fraction had published papers that described them thoroughly (all of which are referenced in this report). For the most part, however, the datasets presented little to no information on data collection schemes. There is also a lack of information on missing data, errors and other limitations. This lack of proper documentation can become a problem for the assessment of data transparency and reliability.

On the positive side, we believe that valuable data have been found while conducting this review. This raises opportunities for new research in the area of AI applied to the railway sector. First, when implementing an open CC dataset, the data can be published together with the research, enhancing its impact and credibility. Data could also be complemented or enhanced using different techniques and made available to the public with the corresponding documentation. At the same time, it would allow to freely compare different models on the same datasets. Second, we observe there is the large amount of government data that is available from most Western countries. Even though most government data are mere statistical information, it can be collected over time to create large datasets using data fusion algorithms. These could be complemented with statistical data and reports made available by governments and rail operating companies.

## 8. Conclusions

Based on previous literature reviews, we have analyzed publicly available datasets across seven railway subdomains for AI applications. We have used multiple portals to collect the public datasets including general databases like European Data Portal and Data.gov, AI-focused databases like Zenodo and FigShare and recent data-focused journals like Data in Brief. The data types have been classified as numeric, image, label and other.

We believe that with the public data available today, some railway problems could already be approached with AI-oriented solutions. For instance, the domain of Traffic Planning and Management counts with a good number of public datasets, in total 28 dataset, and vast availability of data harvesting sources, such as APIs, GTFS, RSS Feeds, for data collection. This data can be used for developing traffic predictions, and also timetabling and real-time rescheduling models and approaches. There are also rather

complete and readily available datasets that can be used to research problems in the domain of Maintenance and Inspection, in total 28 datasets, which are mainly related to maintenance and inspection of railway tracks. Maintenance data for rolling stock, railway ballast and catenary system and electrical equipment are also present but to a lesser extent. This would be used for better health monitoring and predicting failures of infrastructure and rolling stock to minimize the disruption impacts on railway traffic. Other railway domains do not present as many openly available datasets: Safety and Security includes 10 public datasets and Passenger Mobility—8 public datasets. Finally, no datasets could be found for Autonomous Train Driving, Transport Policy or Revenue Management. Thus, generating the first related public datasets could contribute greatly to boosting research in these domains. In all railway domains analyzed in this paper, the most common type of data is numerical data. We believe this is the easiest type of data to obtain, we have observed different data collection methods including mostly onboard sensors, and other IoT devices such as wireless sensor networks. However, only six image datasets could be found. This is a counter-proportional considering the amount of research done involving computer vision models in the railway sector. Also, regarding the datasets quality, it often tend to be rather unknown; and limited or no proper documentation is sometimes available. On the positive side, the study revealed that there are publicly available data maintained by government organizations and TOCs that can be of great use to support AI-based models such as infrastructure network and statistical data, and certified APIs. In particular, these data coming from official public services and are usually well-curated real-time data that can greatly contribute to the accuracy of AI models. Moreover, the increasing number and sophistication of IoT devices—along with the decreasing manufacturing costs—present a promising outlook for the collection of raw data in the railway sector. There are different AI techniques related to Natural Language Processing (NLP) that could be utilized to process any unstructured data collected.

We recognize several promising research directions. First, the formation of a unified database of AI-oriented railway data would be beneficial, we believe that a centralized database of railway-specific datasets would greatly contribute to the conducting research in this area. Second, the further investigation of the quality of the existing datasets would be required to understand its size, quality, and applicability in greater details. Third, the collection of new high-quality data that can be made available for public use and research by active researchers and data and problem owners. Lastly, to guarantee the quality of new datasets, publications in peer-reviewed journals like *Data in Brief* and *IEEE DataPort*, and also online databases like Zenodo and FigShare shall be encouraged. We believe that these new developments would lead towards faster uptake and more diverse developments of AI applications in railway systems.

**Author Contributions:** Conceptualization, M.J.P., F.F., V.V. and N.B.; investigation, M.J.P.; data curation, M.J.P.; writing—original draft preparation, M.J.P.; writing—review and editing, M.J.P., N.B.; supervision, F.F., V.V. and N.B.; revision, M.J.P., F.F., V.V. and N.B.; funding acquisition, F.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work has been partially supported by the research project RAILS (Roadmaps for A.I. integration in the rail Sector). RAILS has received funding from the Shift2Rail Joint Undertaking (JU) under grant agreement No 881782. The JU receives support from the European Union's Horizon 2020 research and innovation program and the Shift2Rail JU members other than the Union.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. List of Datasets

Table A1 presents the list of datasets that have been examined for the review. The datasets are ordered based on the date they were last updated (descending order) and then alphabetically based on their name/title (ascending order).

**Table A1.** List of all 62 datasets consulted for the review.

Ref.	Name/Title	Description	Data Type(s)	Railway Application	Last Updated
[79]	Cityscapes	Video sequences from city scenes in different times and weather conditions. Includes high-quality pixel-level annotations. Includes a class specific to railway vehicles	Image data and label data	Situational intelligence, safety and security	2020
[50]	Image data of spark erosion-ProRail	Images of insulation joints from the Netherlands railway network	Image data	Asset detection, predictive maintenance, fault detection and system monitoring	2020-11
[55]	Evaluating railway track support stiffness from trackside measurements in the absence of wheel load data	Deflection data, trackside measurements, geometrical data and speed data	Numerical data	Predictive maintenance and system monitoring	2020-10
[95]	BART Ridership	Bay Area Rapid Transit (BART) hourly ridership broken up by year (starting 2011)	Numerical data	Passenger flow, traffic planning and infrastructure capacity	2020-10
[31]	Trains Held Short	Weekly average number of trains held short per day	Numerical data and label data	Traffic planning and traffic management	2020-10
[70]	Dataset of measured and commented pantograph electric arcs in DC railways	Digitized sampled data using a data acquisition system located onboard and connected to voltage and current sensors	Numerical data	Maintenance	2020-08
[99]	Predict Train Occupancy Time Series	Monthly data on train occupancy from 1999 to 2011	Numerical data	Passenger flow, traffic planning and infrastructure capacity	2020-08
[46]	Finding railway fasteners in image data-ProRail	Images of fasteners from the Netherlands railway network	Image data	Asset detection, predictive maintenance, fault detection and system monitoring	2020-07
[89]	Experiments on granular flow behavior and deposit characteristics: implications for rock avalanche kinematics	Experimental data on rock avalanche dynamics	Numerical data	Safety and accident prevention	2020-07
[72]	Data sets of measured pantograph voltage and current of European AC railways	Digitized, sampled data using various data recorders located onboard and connected to voltage and current sensors	Numerical data	Maintenance and system monitoring	2020-06
[44]	Bearing Database	Induced failure test data on rolling elements of a spherical roller bearing	Numerical data	Fault detection and predictive maintenance	2020-06
[66]	Cyclic friction tests of ballast stones interfaces under varying vertical load	3D scanner data of angular-tip stones before and after the friction tests	Other data	Maintenance	2020-06

Table A1. Cont.

Ref.	Name/Title	Description	Data Type(s)	Railway Application	Last Updated
[27]	Four small cases for the fairness problem of train timetabling	Data on four small case scenarios for the timetable scheduling fairness problem	Other data	Timetable scheduling	2020-06
[20]	NJ Transit + Amtrak (NEC) Rail Performance	Monthly train trip performance on the NJ Transit rail network	Numerical data and label data	Timetable scheduling, rescheduling and customer satisfaction	2020-05
[76]	Monitoring data for railway bridge KW51 in Leuven, Belgium, before, during, and after retrofitting	Measurements of acceleration on the bridge structures, strain on the rails, displacement at the bearings and more	Numerical data	Maintenance	2020-04
[85]	3D Point Cloud of a railway slope-MOMIT (Multi-scale Observation and Monitoring of railway Infrastructure Threats)	3D point cloud of a railway trench in Lavancia-Épercy in France	Other data	Maintenance and safety	2020-04
[17]	Train delays in Italy Bologna-Milan	Train delays for the Bologna–Milan railway line	Numerical data and label data	Timetable scheduling and rescheduling	2020-03
[57]	An Analysis of Railway Track Behaviour based on Distributed Optical Fiber Acoustic Sensing	Distributed acoustic sensing (DAS) data, plus statistical data from study	Numerical data	Predictive maintenance and system monitoring	2020-02
[62]	3D scans of two types of railway ballast including shape analysis information	3D scanner data of two types of railway ballast, calcite and Kieselkalk	Other data	Maintenance	2020-02
[75]	Performance of Congestion Control Algorithms on High-speed railway scenario	Performance of Hd-TCP and its comparison to simulated data on high-speed railways	Numerical data	Communication	2020-02
[60]	The DR-Train dataset: dynamic responses, GPS positions and environmental conditions of two light rail vehicles in Pittsburgh	Acceleration data, GPS positions, environmental conditions and track maintenance schedules for a light rail network	Numerical data and other data	Predictive maintenance, fault detection and system monitoring	2020-01
[19]	Real-world case based on Batong line in Beijing railway network	Real data for train skip-stopping pattern optimization	Numerical data	Traffic management	2020-01
[81]	RailSem19	Set of 8500 images from a rail vehicle perspective	Image data, numerical data and label data	System monitoring and situational intelligence	2019
[92]	Ungulate-train collision database	Over 3500 animal–train collisions and over 10.000 locations provided by Polish State Railways, PKP	Numerical data and label data	Safety and risk assessment	2019-11

Table A1. Cont.

Ref.	Name/Title	Description	Data Type(s)	Railway Application	Last Updated
[100]	Train Crowd Density	Records of crowd density on several trains over a span of months	Numerical data	Passenger flow, Infrastructure Capacity	2019-11
[16]	SEPTA-Regional Rail	Performance data on regional trains from Southeastern Pennsylvania Transportation Authority	Numerical data, label data	Timetable scheduling, rescheduling	2019-11
[49]	Image data of insulation joints-ProRail	Images of insulation joints from the Netherlands railway network	Image data	Predictive maintenance, fault detection and system monitoring	2019-10
[73]	2 × 25 kV Railway Feeding System Simulation Database	Electricity data measurements from simulations	Numerical data	Maintenance	2019-08
[28]	OpenTrack simulation model files and output dataset for a Copenhagen suburban railway	Simulation data for the evaluation of performance and delays of a suburban railway line	Numerical data	Timetable scheduling and traffic management	2019-08
[38]	Effect of train speed and track geometry on the ride comfort of high-speed railways based on ISO 2631-1	Effects of train speed and track geometry on ride comfort	Numerical data	Predictive maintenance and fault detection	2019-07
[98]	Indian Metro Data	Prediction of future traffic	Numerical data and label data	Passenger flow and traffic management	2019-07
[97]	DBAHN Travel Captures	Data captured from trains and travels in different stations in Germany	Numerical data	Timetable scheduling and passenger flow	2019-06
[24]	IRCTC-Train Info	Data on details of trains of Indian Railways, including timetables and destinations	Numerical data and label data	Timetable scheduling	2019-06
[32]	Integrated train blocking and shipment path optimization (TBSP)	Operational data for resolving the TBSP problem	Numerical data and other data	Route selection and fleet management	2019-05
[23]	Railway Timetable	Timetable data and train details from Indian Railways	Numerical data and label data	Timetable scheduling	2019-05
[87]	Monitoring and early warning method for a rock fall along railways based on vibration signal characteristics	Train vibration signal and rockfall vibration signals captured by sensors	Numerical data	Predictive maintenance, system monitoring and safety	2019-04
[90]	Analytic Geomagnetic and Geoelectric Fields	Geomagnetic and geoelectric field values generated by analytic calculations	Numerical data	Situational intelligence, safety, risk assessment and predictive maintenance	2019-01
[68]	Data on wind-induced responses of the hanging point for a high-speed railway in China	Measurements from micro-acceleration sensor and a laser displacement meter on the catenary systems of HSR	Label data and numerical data	Maintenance	2018-12

Table A1. Cont.

Ref.	Name/Title	Description	Data Type(s)	Railway Application	Last Updated
[94]	Sleep Patterns of Railroad Dispatchers	Data on the work schedules and sleep patterns of railroad employees	Numerical data and label data	Security, safety and risk assessment	2018-11
[77]	Towards the use of UHPFRC in railway bridges: the rehabilitation of Buna Bridge	Acceleration data corresponding to a roving test during an experiment carried out while refactoring bridge	Numerical data	Maintenance	2018-10
[65]	Compression tests and direct shear test of two types of railway ballast	Measurement data from uniaxial compression tests and direct shear tests conducted on railway ballast	Numerical data	Maintenance	2018-09
[45]	Condition of pantograph slide plates	Images of pantograph slide plates from various rolling stock vehicles	Image data	Predictive maintenance, fault detection and system monitoring	2018-07
[22]	Data analysis and visualization of Indian Railways	Trip information on Indian Railways	Numerical data and label data	Timetable scheduling and traffic planning	2018-07
[91]	Accidents in France from 2005 to 2016	Detailed data on traffic accidents in France from 2005 to 2016	Numerical data and label data	Safety and security	2018-06
[21]	Commuter train timetable	Commuter train service in Stockholm during 2012, including timetables and passenger flow	Numerical data and label data	Timetable scheduling and passenger flow	2018-05
[25]	Predicting Near Term Train Schedule Performance and Delay	Data on train deviations from planned schedules for resolving the re-scheduling problem	Numerical data and label data	Timetable scheduling and train re-scheduling	2018-02
[51]	Automated processing of railway track deflection signals obtained from velocity and acceleration measurements	Modeled and measured data for train passages to classify the range of total and downward deflection from train pass-by records	Numerical data	Predictive maintenance, fault detection and system monitoring	2018-01
[42]	Prediction of rail and bridge noise from concrete railway viaducts using a multi-layer rail fastener model and a wavenumber domain method	Noise data in Hz and dB/m	Numerical data	Predictive maintenance and system monitoring	2017
[53]	Investigating the Influence of Auxiliary Rails on Dynamic Behavior of Railway Transition Zone by a 3D Train-Track Interaction Model	Results from different sensitive analyses (vehicle speed, vehicle load, number of auxiliary rails and railpad stiffness) performed with 3D models	Numerical data	Fault detection and predictive maintenance	2017-12
[47]	Fatigue Assessment Method for pre-stressed concrete sleeper	Calculation of remaining fatigue life of concrete sleeper	Numerical data	Predictive maintenance	2017-11



Table A1. Cont.

Ref.	Name/Title	Description	Data Type(s)	Railway Application	Last Updated
[83]	PETS 2017	Data from on-board surveillance systems for protection of critical assets	Image data	Safety, security and system monitoring	2017-07
[40]	Influence of rail fastener stiffness on railway vehicle interior noise	Interior noise in Hz frequency and dB values of different fasteners at different train speeds. Exterior noise. Vibration spectra of train parts	Numerical data	Predictive maintenance and system monitoring	2017-05
[30]	Experimental dataset for optimizing the freight rail operations	Operational data for the development of mathematical models	Label data and numerical data	Logistics and optimization	2016-12
[96]	SBB CFF FFS-Passenger Frequency	Passenger frequency data from Swiss Federal Railways during 2014	Numerical data	Passenger flow	2016-08
[18]	Trains Express Régionaux: Points d'arrêts et horaires des lignes	Timetables of TER trains in France with stops and timetables	Numerical data and label data	Timetable scheduling	2016-08
[26]	Routing Trains through a Railway Network: Joint optimization on train timetabling and maintenance task scheduling	Operational data for resolving the schedule optimization and maintenance task scheduling problem	Other data	Timetable scheduling and maintenance scheduling	2016-07
[59]	Track geometry analytics	Historical detection readings for three types of track defects: surface, cross level and dip	Numerical data	System monitoring, fault detection and predictive maintenance	2015
[33]	Railroad Hump Yard Block-to-Track Assignment	Operational data for resolving the hump yard classification problem	Other data	Fleet management	2014
[34]	Modeling Railroad Yard Capacity	Supporting files for resolving the hump yard capacity modeling problem	Other data	Fleet management	2013
[35]	Movement Planner Algorithm Design for Dispatching on Multi-Track Territories	Supporting files for resolving the multi-track territories dispatching problem	Other data	Fleet management	2012
[36]	Train Design Optimization Problem	Supporting files for resolving the block-to-train assignment problem	Other data	Fleet management and train routing	2011
[37]	Locomotive Refueling Problem	Operational data for resolving the locomotive refueling problem	Other data	Asset management	2010

References

- Schwab, K. Foreign Affairs, The Fourth Industrial Revolution, What It Means and How to Respond. December 2019. Available online: <https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution> (accessed on 3 February 2020).
- David, B. The Future of Intelligence is Artificial. International Railway Journal (IRJ). Available online: [https://www.railjournal.com/in\\_depth/future-intelligence-artificial](https://www.railjournal.com/in_depth/future-intelligence-artificial) (accessed on 3 February 2020).

3. European Parliamentary Research Service (EPRS), European Parliament. Artificial Intelligence in Transport. Current and Future Developments, Opportunities and Challenges. April 2019. Available online: [https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/635609/EPRS\\_BRI\(2019\)635609\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/635609/EPRS_BRI(2019)635609_EN.pdf) (accessed on 3 February 2020).
4. Innovation and Networks Executive Agency (INEA). Horizon 2020 Funding Areas. European Commission. Available online: <https://ec.europa.eu/inea/en/horizon-2020> (accessed on 3 February 2020).
5. Shift2rail.org, "About". Available online: <https://shift2rail.org/about-shift2rail/> (accessed on 3 February 2020).
6. Nakhaee, M.C.; Hiemstra, D.; Stoelinga, M.; van Noort, M. The Recent Applications of Machine Learning in Rail Track Maintenance: A Survey. In Proceedings of the International Conference on Reliability, Safety, and Security of Railway Systems, Lille, France, 4–6 June 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 91–105. [CrossRef]
7. Thilagavathy, N.; Harene, J.; Sherine, M.; Shanmugasundari, T. Survey on railway wheel defect detection using machine learning. *AutAut Res. J.* **2020**, *11*, 4.
8. Liu, S.; Wang, Q.; Luo, Y. A review of applications of visual inspection technology based on image processing in the railway industry. *Transp. Saf. Environ.* **2019**, *1*, 185–204. [CrossRef]
9. Xie, P.; Li, T.; Liu, J.; Du, S.; Yang, X.; Zhang, J. Urban flow prediction from spatiotemporal data using machine learning: A survey. *Inf. Fusion* **2020**, *59*, 1–12. [CrossRef]
10. Wen, C.; Huang, P.; Li, Z.; Lessan, J.; Fu, L.; Jiang, C.; Xu, X. Train Dispatching Management With Data-Driven Approaches: A Comprehensive Review and Appraisal. *IEEE Access* **2019**, *7*, 114547–114571. [CrossRef]
11. Zhu, L.; Yu, F.R.; Wang, Y.; Ning, B.; Tang, T. Big Data Analytics in Intelligent Transportation Systems: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 383–398. [CrossRef]
12. Ghofrani, F.; He, Q.; Goverde, R.; Liu, X. Recent applications of big data analytics in railway transportation systems: A survey. *Transp. Res. Part C Emerg. Technol.* **2018**, *90*, 226–246. [CrossRef]
13. Wu, Q.; Cole, C.; McSweeney, T. Applications of particle swarm optimization in the railway domain. *Int. J. Rail Transp.* **2016**, *4*, 167–190. [CrossRef]
14. Bešinović, N. Deliverable D1.2: Summary of Existing Relevant Projects and State-of-the-Art of AI Application in Railways, RAILS, Shift2Rail. Available online: [https://rails-project.eu/wp-content/uploads/sites/73/2021/05/RAILS\\_D12\\_v23.pdf](https://rails-project.eu/wp-content/uploads/sites/73/2021/05/RAILS_D12_v23.pdf) (accessed on 5 April 2020).
15. Marrone, S.; De Donato, L.; Vittorini, V.; Nardone, R.; Tang, R.; Besinovic, N.; Flammini, F.; Goverde, R.M.P.; Lin, Z. Findings about the State-of-Practice. Deliverable D1.3 Application Areas (Chapter 5). 2021. Available online: [https://rails-project.eu/wp-content/uploads/sites/73/2021/09/RAILS\\_D1\\_3\\_Application\\_Areas\\_v32.pdf](https://rails-project.eu/wp-content/uploads/sites/73/2021/09/RAILS_D1_3_Application_Areas_v32.pdf) (accessed on 15 August 2021).
16. The Southeastern Pennsylvania Transportation Authority (SEPTA). Regional Rail: Predict Arrival Times of Philadelphia's Regional Trains (Version 1). [dataset]. 2019. Available online: <https://www.kaggle.com/septa/on-time-performance> (accessed on 24 September 2020).
17. Cecaj, A. Train Delays in Italy Bologna-Milan (Version 1). [dataset]. 2020. Available online: <https://www.kaggle.com/alketcecaj/train-delays-in-italy-bolognamilan> (accessed on 24 September 2020).
18. Trains Express Régionaux. Trains Express Régionaux: Points D'arrêts et Horaires des Lignes. [dataset]. 2016. Available online: <https://www.data.gouv.fr/en/datasets/trains-express-regionaux-points-darrets-et-horaires-des-lignes/> (accessed on 24 September 2020).
19. Yinghui, W. Real-World Case Based on Batong Line in Beijing Railway Network. Mirror of Mendeley Data. [dataset]. 2020. Available online: [https://figshare.com/articles/dataset/Real-world\\_case\\_based\\_on\\_Batong\\_line\\_in\\_Beijing\\_railway\\_network/11627880](https://figshare.com/articles/dataset/Real-world_case_based_on_Batong_line_in_Beijing_railway_network/11627880) (accessed on 5 October 2020).
20. Pranav, B. NJ Transit + Amtrak (NEC) Rail Performance (Version 2). [dataset]. 2020. Available online: <https://www.kaggle.com/pranavbadami/nj-transit-amtrak-nec-performance> (accessed on 24 September 2020).
21. Abderrahman, A.A. Commuter Train Timetable: Commuter Train Service in Stockholm 2012 (Version 2). [dataset]. 2012. Available online: <https://www.kaggle.com/abdeaitali/commuter-train-timetable> (accessed on 25 September 2020).
22. Tanima, S. Data Analysis and Visualization of Indian Railways (Version 1). [dataset]. 2018. Available online: <https://www.kaggle.com/tanimasarkhel/data-analysis-and-visualization-of-indian-railways> (accessed on 25 September 2020).
23. Harshit, G. Indian Railways Time Table for Trains Available (Version 1). [dataset]. 2019. Available online: <https://www.kaggle.com/harshit16/indian-railways-time-table-for-trains-available> (accessed on 25 September 2020).
24. Binil, J. IRCTC-TrainInfo (Version 1). [dataset]. 2019. Available online: <https://www.kaggle.com/binilj04/irctctraininfo> (accessed on 26 September 2020).
25. The Institute for Operations Research and the Management Sciences (INFORMS). Predicting Near Term Train Schedule Performance and Delay (Version 1). [dataset]. 2018. Available online: <https://connect.informs.org/railway-applications/new-item3/problem-repository16> (accessed on 26 September 2020).
26. The Institute for Operations Research and the Management Sciences (INFORMS). Routing Trains through a Railway Network: Joint Optimization on Train Timetabling and Maintenance Task Scheduling. [dataset]. 2016. Available online: <https://connect.informs.org/railway-applications/new-item3/problem-repository16> (accessed on 26 September 2020).
27. Yinghui, W. Four Small Cases for the Fairness Problem of Train Timetabling. [dataset]. 2020. Available online: [https://mendeley.figshare.com/articles/dataset/Four\\_small\\_cases\\_for\\_the\\_fairness\\_problem\\_of\\_train\\_timetabling/12402911](https://mendeley.figshare.com/articles/dataset/Four_small_cases_for_the_fairness_problem_of_train_timetabling/12402911) (accessed on 26 September 2020).

28. Harrod, S.; Cerreto, F.; Nielsen, O.A. OpenTrack simulation model files and output dataset for a Copenhagen suburban railway. *Data Brief* **2019**, *25*, 103952. [CrossRef]
29. Harrod, S.; Cerreto, F.; Nielsen, O.A. A closed form railway line delay propagation model. *Transp. Res. Part C Emerg. Technol.* **2019**, *102*, 189–209. [CrossRef]
30. Mahmoud, M.; Erhan, K.; Geoff, K.; Shi, Q.L. Experimental dataset for optimizing the freight rail operations. *Data Brief* **2016**, *9*, 492–500. [CrossRef]
31. Jesse, G.; (Surface Transportation Board). Trains Held Short. [dataset]. 2020. Available online: <https://agtransport.usda.gov/Rail/Trains-Held-Short/iacs-9uck> (accessed on 5 October 2020).
32. The Institute for Operations Research and the Management Sciences (INFORMS) Railway Application Section (RAS). Integrated Train Blocking and Shipment Path Optimization (TBSP) (Version 1). [dataset]. 2019. Available online: <https://connect.informs.org/railway-applications/new-item3/problem-repository16> (accessed on 1 November 2020).
33. The Institute for Operations Research and the Management Sciences (INFORMS) Railway Application Section (RAS). Railroad Hump Yard Block-to-Track Assignment. [dataset]. 2014. Available online: <https://connect.informs.org/railway-applications/new-item3/problem-repository16> (accessed on 1 November 2020).
34. The Institute for Operations Research and the Management Sciences (INFORMS) Railway Application Section (RAS). Modeling Railroad Yard Capacity. [dataset]. 2013. Available online: <https://connect.informs.org/railway-applications/new-item3/problem-repository16> (accessed on 1 November 2020).
35. The Institute for Operations Research and the Management Sciences (INFORMS) Railway Application Section (RAS). Movement Planner Algorithm Design for Dispatching on Multi-Track Territories. [dataset]. 2012. Available online: <https://connect.informs.org/railway-applications/new-item3/problem-repository16> (accessed on 1 November 2020).
36. The Institute for Operations Research and the Management Sciences (INFORMS) Railway Application Section (RAS). Train Design Optimization Problem. [dataset]. 2011. Available online: <https://connect.informs.org/railway-applications/new-item3/problem-repository16> (accessed on 1 November 2020).
37. The Institute for Operations Research and the Management Sciences (INFORMS) Railway Application Section (RAS). Locomotive Refueling Problem. [dataset]. 2010. Available online: <https://connect.informs.org/railway-applications/new-item3/problem-repository16> (accessed on 1 November 2020).
38. Liu, C.; Thompson, D.; Griffin, M.J.; Entezami, M. Dataset for “Effect of Train Speed and Track Geometry on the Ride Comfort Of High-Speed Railways Based on ISO 2631-1”. University of Southampton. [dataset]. 2019. Available online: <https://eprints.soton.ac.uk/432605/> (accessed on 5 October 2020).
39. Liu, C.; Thompson, D.; Griffin, M.J.; Entezami, M. Effect of train speed and track geometry on the ride comfort in high-speed railways based on ISO 2631-1. *Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit* **2019**, *234*, 765–778. [CrossRef]
40. Li, L.; Thompson, D.; Xie, Y.; Zhu, Q.; Luo, Y.; Lei, Z. Dataset for Influence of Rail Fastener Stiffness on Railway Vehicle Interior Noise. University of Southampton. [dataset]. 2017. Available online: <https://eprints.soton.ac.uk/428923/> (accessed on 10 August 2020).
41. Li, L.; Thompson, D.; Xie, Y.; Zhu, Q.; Luo, Y.; Lei, Z. Influence of rail fastener stiffness on railway vehicle interior noise. *Appl. Acoust.* **2018**, *145*, 69–81. [CrossRef]
42. Li, Q.; Thompson, D. Dataset for Paper: Prediction of Rail and Bridge Noise from Concrete Railway Viaducts Using a Multi-Layer Rail Fastener Model and a Wavenumber Domain Method. University of Southampton. [dataset]. 2017. Available online: <https://eprints.soton.ac.uk/411733/> (accessed on 10 August 2020).
43. Li, Q.; Thompson, D. Prediction of rail and bridge noise arising from concrete railway viaducts by using a multilayer rail fastener model and a wavenumber domain method. *Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit* **2017**, *232*, 1326–1346. [CrossRef]
44. César, R.S.-O.; José, M.M.; Juan, D.C.-M.; José, L.; Garcia, B. Bearing Database (Version V1). [dataset]. 2020. Available online: <https://zenodo.org/record/3898942> (accessed on 10 August 2020).
45. Gehrig, U. Condition of Pantograph Slide Plates: Images From Pantograph Slide Plates of Various Rolling Stock Vehicles (Version 5). [dataset]. 2018. Available online: <https://www.kaggle.com/gehrig/pantograph> (accessed on 2 September 2020).
46. van Hees, O. Finding Railway Fasteners in Image Data—ProRail (Version 4). [dataset]. 2020. Available online: <https://www.kaggle.com/oscarvanhees/finding-railway-fasteners-in-image-data-prorail> (accessed on 20 September 2020).
47. Ruilin, Y.; Dan, L.; Chayut, N.; Rims, J.; Sakdirat, K. Fatigue Assessment Method for Pre-Stressed Concrete Sleeper (Version 2). [dataset]. 2017. Available online: [https://zenodo.org/record/1155711#.YI6t\\_ej7Stp](https://zenodo.org/record/1155711#.YI6t_ej7Stp) (accessed on 15 August 2020).
48. You, R.; Li, D.; Ngamkhanong, C.; Janeliukstis, R.; Kaewunruen, S. Fatigue Life Assessment Method for Prestressed Concrete Sleepers. *Front. Built Environ.* **2017**, *3*, 68. [CrossRef]
49. van Hees, O. Image Data of Insulation—ProRail: Image Recognition Used for Asset Detection. [dataset]. 2019. Available online: <https://www.kaggle.com/oscarvanhees/insulation-joint-training-set-prorail> (accessed on 9 November 2020).
50. van Hees, O. Image Data of Spark Erosion—ProRail (Version 3). [dataset]. 2020. Available online: <https://www.kaggle.com/oscarvanhees/image-data-of-spark-erosion-prorail> (accessed on 10 January 2021).
51. Milne, D. Automated Processing of Railway Track Deflection Signals Obtained from Velocity and Acceleration Measurements. [dataset]. 2018. Available online: <http://eprints.soton.ac.uk/id/eprint/419011> (accessed on 15 January 2021).
52. Milne, D.; Pen, L.L.; Thompson, D.; Powrie, W. Automated processing of railway track deflection signals obtained from velocity and acceleration measurements. *Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit* **2018**, *232*, 2097–2110. [CrossRef]

53. Heydari-Noghabi, H.; Varandas, J.N.; Esmaeili, M.; Zakeri, J. Investigating the Influence of Auxiliary Rails on Dynamic Behavior of Railway Transition Zone by a 3D Train—Track Interaction Model. [dataset]. 2017. Available online: [https://figshare.com/articles/dataset/Investigating\\_the\\_Influence\\_of\\_Auxiliary\\_Rails\\_on\\_Dynamic\\_Behavior\\_of\\_Railway\\_Transition\\_Zone\\_by\\_a\\_3D\\_Train-Track\\_Interaction\\_Model/5734317](https://figshare.com/articles/dataset/Investigating_the_Influence_of_Auxiliary_Rails_on_Dynamic_Behavior_of_Railway_Transition_Zone_by_a_3D_Train-Track_Interaction_Model/5734317) (accessed on 15 August 2020).
54. Heydari-Noghabi, H.; Varandas, J.N.; Esmaeili, M.; Zakeri, J. Investigating the Influence of Auxiliary Rails on Dynamic Behavior of Railway Transition Zone by a 3D Train-Track Interaction Model. *Lat. Am. J. Solids Struct.* **2017**, *14*, 2000–2018. [CrossRef]
55. Le Pen, L.; Milne, D.; Thompson, D.; Powrie, W. *Evaluating Railway Track Support Stiffness from Trackside Measurements in the Absence of Wheel Load Data*; University of Southampton: Southampton, UK, 2016. [CrossRef]
56. Le Pen, L.; Milne, D.; Thompson, D.; Powrie, W. Evaluating railway track support stiffness from trackside measurements in the absence of wheel load data. *Can. Geotech. J.* **2016**, *53*, 1156–1166. [CrossRef]
57. Milne, D.; le Pen, L.; Watson, G.; Masoudi, A. Data for: An Analysis of Railway Track Behaviour based on Distributed Optical Fibre Acoustic Sensing (Version 1). University of Southampton. [dataset]. 2020. Available online: <https://eprints.soton.ac.uk/438063/> (accessed on 3 December 2020).
58. Milne, D.; Masoudi, A.; Ferro, E.; Watson, G.; Le Pen, L. An analysis of railway track behaviour based on distributed optical fibre acoustic sensing. *Mech. Syst. Signal Process.* **2020**, *142*, 106769. [CrossRef]
59. The Institute for Operations Research and the Management Sciences (INFORMS) Railway Application Section (RAS). Track Geometry Analytics. [dataset]. 2015. Available online: <https://connect.informs.org/railway-applications/new-item3/problem-repository16> (accessed on 1 November 2020).
60. Liu, J.; Chen, S.; Lederman, G.; Kramer, D.B.; Noh, H.Y.; Bielak, J.; Berges, M. The DR-Train dataset: Dynamic Responses, GPS Positions and Environmental Conditions Of Two Light Rail Vehicles in Pittsburgh (Version 1.0). [dataset]. 2018. Available online: <https://zenodo.org/record/1432702#.YLPmY6j7Sto> (accessed on 29 July 2020).
61. Liu, J.; Chen, S.; Lederman, G.; Kramer, D.B.; Noh, H.Y.; Bielak, J.H.G., Jr.; Kovačević, J.; Bergés, M. Dynamic responses, GPS positions and environmental conditions of two light rail vehicles in Pittsburgh. *Sci. Data* **2019**, *6*, 146. [CrossRef] [PubMed]
62. Suhr, B.; Six, K.; Skipper, W.A.; Lewis, R. 3D Scans of Two Types of Railway Ballast Including Shape Analysis Information (Version 1). [dataset]. 2020. Available online: <https://zenodo.org/record/3689592> (accessed on 1 December 2020).
63. Suhr, B.; Skipper, W.A.; Lewis, R.; Six, K. Shape analysis of railway ballast stones: Curvature-based calculation of particle angularity. *Sci. Rep.* **2020**, *10*, 6045. [CrossRef]
64. Suhr, B.; Six, K. Simple particle shapes for DEM simulations of railway ballast: Influence of shape descriptors on packing behaviour. *Granul. Matter* **2020**, *22*, 43. [CrossRef] [PubMed]
65. Suhr, B.; Six, K. Compression Tests and Direct Shear Test of Two Types of Railway Ballast (Version 1). [dataset]. 2018. Available online: <https://zenodo.org/record/1423742#.YI6RR-j7Stp> (accessed on 29 July 2020).
66. Suhr, B.; Butcher, T.A.; Lewis, R.; Six, K. Cyclic Friction Tests of Ballast Stones Interfaces Under Varying Vertical Load (Version 1). [dataset]. 2020. Available online: <https://zenodo.org/record/3893842> (accessed on 22 November 2020).
67. Suhr, B.; Butcher, T.; Lewis, R.; Six, K. Friction and wear in railway ballast stone interfaces. *Tribol. Int.* **2020**, *151*, 106498. [CrossRef]
68. Xie, Q.; Zhi, X. Data on wind-induced responses of the hanging point for a high-speed railway in China. *Data Brief* **2018**, *21*, 2259–2261. [CrossRef] [PubMed]
69. Xie, Q.; Zhi, X. Wind tunnel test of an aeroelastic model of a catenary system for a high-speed railway in China. *J. Wind. Eng. Ind. Aerodyn.* **2018**, *184*, 23–33. [CrossRef]
70. Signorino, D.; Giordano, D.; Mariscotti, A.; Gallo, D.; Femine, A.D.; Balic, F.; Quintana, J.; Donadio, L.; Biancucci, A. Dataset of measured and commented pantograph electric arcs in DC railways. *Data Brief* **2020**, *31*, 105978. [CrossRef]
71. MyRailS. MyRailS: Accurate Measurements for Energy Efficiency in European Railway and Subway Systems. Available online: <https://myrails.it/> (accessed on 1 October 2020).
72. Mariscotti, A. Data sets of measured pantograph voltage and current of European AC railways. *Data Brief* **2020**, *30*, 105477. [CrossRef] [PubMed]
73. Arboleya, P.; Mohamed, B.; El-Sayed, I.; Gonzalez-Moran, C. 2 × 25 kv Railway Feeding System Simulation Database, IEEE Dataport. [dataset]. 2019. Available online: <https://ieee-dataport.org/documents/2x25kv-railway-feeding-system-simulation-database> (accessed on 10 September 2020).
74. Mohamed, B.; Arboleya, P.; ElSayed, I.; Gonzalez-Moran, C.; El-Sayed, I. High-Speed 2 × 25 kV Traction System Model and Solver for Extensive Network Simulations. *IEEE Trans. Power Syst.* **2019**, *34*, 3837–3847. [CrossRef]
75. Yuan, Z. Performance of Congestion Control Algorithms on High-Speed Railway Scenairo (Version 1), IEEE Dataport. [dataset]. 2020. Available online: <https://ieee-dataport.org/documents/performance-congestion-control-algorithms-high-speed-railway-scenairo> (accessed on 12 December 2020).
76. Maes, K.; Lombaert, G. Monitoring Data for Railway Bridge KW51 In Leuven, Belgium, Before, During, and after Retrofitting (Version 1.0). [dataset]. 2020. Available online: <https://zenodo.org/record/3745914> (accessed on 10 December 2020).
77. Martin-Sanz, H.; Tatsis, K.; Stipanovic, I.; Damjanovic, D.; Sanja, A.; Brühwiler, E.; Chatzi, E. Towards the use of UHPFRC in railway bridges: The rehabilitation of Buna Bridge (Version 1). [dataset]. 2008. Available online: <https://zenodo.org/record/2574457#.YI5hGOj7Stq> (accessed on 8 August 2020).

78. Martín-Sanz, H.; Tatsis, K.; Chatzi, E.; Brühwiler, E.; Stipanovic, I.; Mandic, A.; Damjanovic, D.; Sanja, A. Towards the use of UHPFRC in railway bridges: The rehabilitation of Buna Bridge. In Proceedings of the 5th International Symposium on Life-Cycle Civil Engineering (IALCCE 2018), Lake Como, Italy, 11–14 June 2018; 14 June 2018.
79. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. [dataset]. 2016. Available online: <https://www.cityscapes-dataset.com> (accessed on 10 August 2020).
80. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. 2016. Available online: <https://www.cityscapes-dataset.com> (accessed on 10 August 2020).
81. Zendel, O.; Murschitz, M.; Zeilinger, M.; Steininger, D.; Abbasi, S.; Beleznai, C. RailSem19: A Dataset for Semantic Rail Scene Understanding (Version 1). [dataset]. 2019. Available online: <https://wilddash.cc/railsem19> (accessed on 25 August 2020).
82. Zendel, O.; Murschitz, M.; Zeilinger, M.; Steininger, D.; Abbasi, S.; Beleznai, C. A Dataset for Semantic Rail Scene Understanding. Conference on Computer Vision and Pattern Recognition (CVPR) Workshop, Long Beach, CA, USA, 16–20 June 2019.
83. Patino, L.; Nawaz, T.; Cane, T.; Ferryman, J. PETS 2017. [dataset]. 2017. Available online: <https://doi.org/10.1109/CVPRW.2017.264> (accessed on 23 August 2020).
84. Patino, L.; Nawaz, T.; Cane, T.; Ferryman, J. PETS 2017: Dataset and Challenge. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 2126–2132.
85. Riquelme, J.L.A.; Ruffo, M.; Tomás, R.; Riquelme, A.; Pagán, J.I.; Cano, M.; Pastor, J.L. 3D Point Cloud of a Railway Slope—MOMIT (Multi-Scale Observation And Monitoring of Railway Infrastructure Threats) EU Project—H2020-EU.3.4.8.3.—Grant Agreement ID: 777630. [dataset]. 2020. Available online: <https://zenodo.org/record/3777996> (accessed on 2 December 2020).
86. MOMIT Project Consortium. MOMIT: Multi-scale Observation and Monitoring of railway Infrastructure Threats. Available online: <https://www.momit-project.eu/> (accessed on 1 October 2020).
87. Yan, Y.; Li, T.; Liu, J.; Wang, W.; Su, Q. Monitoring and Early Warning Method For A Rock Fall Along Railways Based On Vibration Signal Characteristics. [dataset]. 2019. Available online: <https://www.nature.com/articles/s41598-019-43146-1> (accessed on 1 October 2020).
88. Yan, Y.; Li, T.; Liu, J.; Wang, W.; Su, Q. Monitoring and early warning method for a rockfall along railways based on vibration signal characteristics. *Sci. Rep.* **2019**, *9*, 6606. [CrossRef] [PubMed]
89. Li, K.; Wang, Y.; Lin, Q.; Cheng, Q.; Wu, Y. Experiments on Granular Flow Behavior and Deposit Characteristics: Implications For Rock Avalanche Kinematics (Version 3). [dataset]. 2020. Available online: <https://zenodo.org/record/3930161#.YHM1Qui7Stp> (accessed on 1 November 2020).
90. Boteler, D.; Pirjola, R.; Marti, L. Analytic Geomagnetic and Geoelectric Fields, IEEE Dataport. [dataset]. 2019. Available online: <https://ieee-dataport.org/open-access/analytic-geomagnetic-and-geoelectric-fields> (accessed on 8 August 2020).
91. Mimi, A.L. Accidents in France from 2005 to 2016 (Version 2). [dataset]. 2018. Available online: <https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016> (accessed on 8 August 2020).
92. Jasińska, D.; Żmihorski, M.; Krauze-Gryz, D.; Kotowska, D.; Werka, J.; Piotrowska, D.; Pärt, T. Data From: Linking Habitat Composition, Local Population Densities and Traffic Characteristics to Spatial Patterns of Ungulate-Train Collisions. [dataset]. 2019. Available online: <https://datadryad.org/stash/dataset/doi:10.5061/dryad.870t013> (accessed on 8 August 2020).
93. Jasińska, K.D.; Żmihorski, M.; Krauze-Gryz, D.; Kotowska, D.; Werka, J.; Piotrowska, D.; Pärt, T. Linking habitat composition, local population densities and traffic characteristics to spatial patterns of ungulate-train collisions. *J. Appl. Ecol.* **2019**, *56*, 2630–2640. [CrossRef]
94. Minasyan, N. Sleep Patterns of Railroad Dispatchers: How well Railroad Dispatchers Sleep (Version 1). [dataset]. 2018. Available online: <https://www.kaggle.com/nairaminasyan/sleep-patterns> (accessed on 8 August 2020).
95. Geislinger, V. BART Ridership (Version 6). [dataset]. 2020. Available online: <https://www.kaggle.com/mrgeislinger/barridership> (accessed on 20 August 2020).
96. Bellanger, A. SBB CFF FFS—Passenger Frequency. [dataset]. 2016. Available online: <https://data.world/antoinebell/sbb-passengerfrequency> (accessed on 20 August 2020).
97. Mengibar, C. D BAHN Travels Captures: Data Captured from Trains And Travels in Different Station Of Germany (Version 1). [dataset]. 2019. Available online: <https://www.kaggle.com/chemamengibar/dbahn-travels-captures> (accessed on 10 August 2020).
98. Ansari, U. Indian Metro Data: Prediction of the Future Traffic (Version 1). [dataset]. 2019. Available online: <https://www.kaggle.com/umairnsr87/indian-metro-data> (accessed on 8 August 2020).
99. Reddy, R. Predict Train Occupancy Time Series (Version 1). [dataset]. 2020. Available online: <https://www.kaggle.com/gajjadarahul/predict-train-occupancy-time-series> (accessed on 3 October 2020).
100. Tyagi, A. Train Crowd Density: Details of Several Trains Along with Target Variable Being Crowd Density (Version 1). [dataset]. 2019. Available online: <https://www.kaggle.com/akashtyagi08/trainnn> (accessed on 8 August 2020).
101. Silva, F.B.E.; Forzieri, G.; Herrera, M.A.M.; Bianchi, A.; LaValle, C.; Feyen, L. HARCIEU, a harmonized gridded dataset of critical infrastructures in Europe for large-scale risk assessments. *Sci. Data* **2019**, *6*, 126. [CrossRef]