

Student thesis series INES nr 560

# Digitalization in the rail industry: Localizing damaged cargo wagons using spatial operations and big data

**Julia Hellner**

---

2021  
Department of  
Physical Geography and Ecosystem Science  
Lund University  
Sölvegatan 12  
S-223 62 Lund  
Sweden



Julia Hellner (2021).

***Digitalization in the rail industry:  
Localizing damaged cargo wagons using spatial operations and big data***

Master degree thesis, 30 credits in Geomatics

Department of Physical Geography and Ecosystem Science, Lund University

Level: Master of Science (MSc)

Course duration: *January 2021 until September 2021*

Disclaimer

This document describes work undertaken as part of a program of study at the University of Lund. All views and opinions expressed herein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

Digitalization in the rail industry:  
Localizing damaged cargo wagons using spatial  
operations and big data

---

Julia Hellner

Master thesis, 30 credits, in Geomatics

**Supervisors:**

Micael Runnström

Department of Physical Geography and Ecosystem Science, Lund University

Alexander Weiß

DB Cargo AG

Daniel Rost

DB Cargo AG

**Exam committee:**

Jonas Ardö

Department of Physical Geography and Ecosystem Science, Lund University



## Acknowledgements

I want to thank the people that supported me during my thesis journey. Micael Runnström, Daniel Rost, and Alexander Weiß constantly guided, and challenged me to improve my work and scientific approach. I want to thank Max Mangold for his technical feedback and honest interest. Further, I am grateful for Amir Jaber and my sister Isabell for proof-reading and giving feedback from a different perspective. I also want to thank DB Cargo, and especially Gerrit Koch to Krax and the Wagon Intelligence project for giving me the opportunity to gain practical experience in the rail industry and participate in innovative use cases. Lastly, I want to thank my family and friends for motivating me, distracting me, and giving me the opportunity to refuel.

## Abstract

Climate change and the sustainability challenges of the 21<sup>st</sup> century require the reduction of greenhouse gas emissions in all sectors. Accordingly, political objectives urge for a shift from road to rail in logistics. This drives innovation, increased efficiency, and customer friendliness in the rail industry. These advancements are strongly needed as investments have been neglected in the past and the technological development lags behind the standard in the road freight sector. The use of Global Navigation Satellite System (GNSS) wagon data to allow tracking, optimized fleet management and maintenance is one effort to compete with road freight. This thesis proposes an algorithm that can be used to gain information on the repair status of a wagon while abroad. As cooperation and data exchange between railway companies are not well developed, delays in repair and disposition occur regularly. The created algorithm estimates the day a damaged wagon was back in operation, information that is not available today for most damage cases. In the process, geofences for cargo rail maintenance facilities in Europe are created automatically. A methodology using existing address, infrastructure, and GNSS big data in combination with spatial analysis and clustering tools to identify insufficiently referenced geographic objects is proposed. Data from the largest rail operator in Europe, Deutsche Bahn AG, is used in the analysis.

The created algorithm computes the day a wagon was back in operation in over 86 % of damage cases. The result deviates on average four days from the actual day of the operations release. Although it was not possible to conclusively rate the quality of all geofences, the data and feedback from maintenance employees indicate that most of them are reliable. Faulty addresses, too small geofences, mobile maintenance, and damage protocol errors were the main reasons for unsuccessful operations release computations.

**Keywords:** Big data, clustering, digitization, geofence, geomatics, GNSS, maintenance, rail, spatial analysis

---

## Table of contents

Figures .....	VIII
Tables .....	IX
Abbreviations .....	IX
1. Introduction.....	1
1.1 The organization of cargo rail in Europe.....	1
1.2 Challenges of cargo rail in Europe .....	2
1.3 Past research and knowledge gap .....	3
1.4 Research objective .....	4
2. Background and related work .....	6
2.1 Railway industry in Europe .....	6
2.2 Big data and data-driven research .....	8
2.3 Geographical Information Systems .....	9
3. Methodology .....	13
3.1 Input data .....	13
3.2 Data exploration and preprocessing .....	18
3.3 Creation of maintenance geofences .....	19
3.4 Development of algorithm for geofence event detection .....	25
3.5 Evaluation and validation .....	28
4. Results.....	30
4.1 Geofence creation .....	30
4.2 Algorithm development.....	33
5. Discussion .....	38
5.1 Interpretation of results.....	38
5.2 Reflection on validity, objectivity, and reliability of the results .....	39
5.3 Results within research context .....	41
6. Conclusion .....	43
References .....	44
Appendix .....	48

## Figures

Figure 1: Visualization of functionality of DBSCAN clustering.....	11
Figure 2: Behavior of different point clustering algorithms for point dataset examples and their computation time.....	12
Figure 3: Study areas with national rail network .....	14
Figure 4: Number of wagons owned by DB Cargo that are equipped with telematics units 2018 to 2021.....	15
Figure 5: Steps performed to preprocess OSM buildings data and filter out relevant buildings. ....	18
Figure 6: General process used to create infrastructure geofences from maintenance facility addresses and OSM data. ....	20
Figure 7: Example for process of infrastructure geofence computation for SweMaint AB in Göteborg.....	20
Figure 8: Process of TGF creation .....	23
Figure 9: Cases for combination of IGFs and TGFs and process for each case.....	25
Figure 10: Most important steps performed in algorithm to derive geofence entries and exits and substitute operations release. ....	26
Figure 11: Examples of geofence categories.....	29
Figure 12: Maintenance facilities in Europe identified through DB and VPI address data. ....	30
Figure 13: Locations in Sweden and Italy where the created geofences are situated .....	31
Figure 14: Examples of final geofences in Italy and Sweden .....	32
Figure 15: Example for feedback from Duroc Rail AB on maintenance location in Luleå.....	32
Figure 16: Overview of effectiveness of algorithm in calculating substitute operations releases (SOR). ....	33
Figure 17: Distribution of deviation in SOR date from actual OR date in days for damage cases from 01.04.21 to 30.06.2021 in Italy and Sweden .....	34
Figure 18: Distribution of deviation in SOR date from actual OR date in days for damage cases disaggregated according to quality indicator .....	35
Figure 19: Distribution of deviation in SOR date from actual OR date in days for damage cases disaggregated according to the category of the last relevant geofence exited .....	37

## Tables

Table 1: Relevant information provided by the telematics dataset for each data point. ....	15
Table 2: Relevant information included in downloads from WADIS.....	16
Table 3: Definition of quality indicators $Q_{GF}$ and $Q_{SOR}$ .....	29
Table 4: Statistical values for boxplots in Figure 17, giving the SOR deviation for the entire dataset and disaggregated for each Italy and Sweden .....	35
Table 5: Statistical values for boxplots in Figure 18, SOR deviation disaggregated according to quality indicator $Q$ .....	36
Table 6: Statistical values for boxplots in Figure 19, SOR deviation disaggregated according to geofence category.....	37

## Abbreviations

DB	Deutsche Bahn
DBSCAN	Density-based spatial clustering of application with noise
DSB	Danske Statsbanar
EC	European Commission
EU	European Union
GCU	General Contract of Use for Wagons
GIS	Geographic Information System
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GSM	Global System for Mobile Communications
HDBSCAN	Hierarchical DBSCAN
HDFS	Hadoop Distributed File System
IGF	Infrastructure geofence
NFC	Near-field communication
OR	Operations release
OSM	OpenStreetMap
OTIF	Intergovernmental Organization for the International Carriage by Rail
RC	Rail company
RFID	Radio-frequency identification
SEM	Standard error of the mean
SJ	Statens Järnvägar
SOR	Substitute operations release
TGF	Telemetry geofence
UHF	Ultra high frequency
VPI	Association of Freight Wagon Holder in Germany
WADIS	Wagon disposition and information system



# 1. Introduction

As globalization continuously strengthens the economic ties between countries, transportation of goods becomes more and more crucial for functioning societies. At the same time, the distance goods cover during their lifecycle increases. More than 75 % of goods in the European Union (EU) are conveyed on roads (Eurostat 2020). Freight rail has a modal share of only 18 %. In several countries, e.g. Sweden, the UK, and Germany, the amount of goods transported with trucks has increased over the past years (Eurostat 2020). This is a concerning trend as road cargo transportation causes more than six times the greenhouse gas emissions per ton-kilometer than rail freight (Umweltbundesamt 2019). Currently, freight emissions account for 30 % of global transportation emissions and 7 % of total global emissions (International Transport Forum 2015). As countries aim to decrease carbon emissions to limit global warming, a shift to more sustainable cargo transportation is necessary. Several initiatives with this aim have been started on the national and European level (e.g. Shift2Rail by the EU, Master Plan Rail Transport by Germany, National Transport Plan 2018-2029 by Sweden). However, the change from road to rail has been slow due to the complexity of the railway industry and the neglect of investments into new technology, customer friendliness, international interoperability, and efficiency in the past (Ramirez et al. 2020).

## 1.1 The organization of cargo rail in Europe

To comprehend the challenges cargo rail faces today, the structure of the rail industry in Europe needs to be understood. The railroad has a long tradition in Europe. At the end of the 19<sup>th</sup> century all major European countries had a national railway network (Millward 2005). They were a key driver for the industrialization as they enabled a more efficient transport of raw materials like coal and metal ores (Sieferle 2008). Economic hardship in combination with a public interest in keeping up operations drove most rail companies (RC) into state ownership in the first half of the 20<sup>th</sup> century (Millward 2005). As the public investment focus shifted towards road infrastructure in the second half of the 20<sup>th</sup> century, the extent and quality of the railway network declined (Roth and Divall 2015). Efforts by the European Commission (EC) to again increase the modal share of the railway induced improvements in interoperability and introduced market liberalizations in the early 2000s (European Commission 2001). Since 2007 any licensed RC can offer national and international freight services throughout the EU (European Commission 2004). The track infrastructure and large RCs are still mostly in public ownership, however. Prominent examples for state-owned RCs in Europe are Deutsche Bahn AG (DB) in Germany, Danske Statsbaner (DSB) in Denmark, and Statens Järnvägar (SJ) in Sweden.

In the early stages, national interests and varying track gauges only lead to few cross-border routes. With the progression of the industrialization however, economic ties and the division of labor required a European rail network (Roth and Dinshobl 2008). RCs started working together

to convey goods across the continent. Insufficient interoperability in technical standards, safety, language, and operating regulations led to the establishment of a system in which the national cargo RCs transport wagons only on their national track network (Rail Cargo Group 2019). The wagons containing the goods are consigned to the according RC at the border. This structure is still in place today. Consequently, if goods need to be transported from Hamburg to Prague for example, DB Cargo transfers the respective wagons to ČD Cargo at the Czech border.

Although this system has advantages, it also creates difficulties as a close collaboration between the RCs is needed to enable an efficient operation. This collaboration includes the exchange of information about the route, the destination, and the current position of wagons. Also, an agreement about financial compensation for the transportation and the corresponding billing infrastructure needs to be established. Lastly, a plan of action to handle special circumstances like the prolonged stay of a wagon at a client site, or the damage of a wagon needs to be agreed upon. The General Contract of Use for Wagons (GCU) regulates damaged wagon handling in Europe and is based on recommendations by the Intergovernmental Organization for International Carriage by Rail (OTIF) (GCU 2020). Since 2007, all major RCs, maintenance providers, and freight rail users have joined the GCU, now amounting to over 700 signatories (GCU 2020). According to the GCU, an RC that detects a damage on a conveyed wagon is obliged to restore the operability of the wagon. The following information needs to be sent to the wagon owner by the operating RC:

- Wagon number
- Type of damage
- Date the damage occurred
- Place the damage occurred
- Requests for spare parts if needed
- For larger repairs: cost estimate for the repair
- Date the repair was finished (operations release (OR))
- Invoice if the wagon owner is liable for the damage (e.g. for wear damages)

If the damage was caused by the foreign RC or a client, the wagon owner can demand financial compensation for the damage and the days the wagon could not be used (GCU 2020).

## 1.2 Challenges of cargo rail in Europe

In practice, the GCU process often does not work as described. According to DB Cargo, damage reports and ORs are regularly sent weeks after the repair has been completed, if at all. Often, wagons are not directly transferred to a maintenance facility but stay idle on a railway siding, sometimes for months. As RCs can keep tens of thousands of wagons, it takes time to track the status of each of them. DB Cargo for example operates over 90,000 wagons (DB Cargo AG 2021b) and loses track of single wagons at times, especially if they are abroad. The amount of available information abroad is much lower than when a wagon is on its “home” rail network.

This is caused by an insufficient interconnection of the RC's IT systems. Even today, several phone calls are sometimes necessary to determine the approximate location of a specific wagon, needed to tell a client where their goods are.

The freight rail industry lags behind when it comes to technological advancements (Berrios Villalba 2020). The long lifecycle of rail infrastructure impedes adapting new and evolving technologies. At the same time, transportation costs for customers need to remain low to stay competitive with road freight transport (Berrios Villalba 2020). Consequently, the technical features of freight wagons are still very basic. The wagons are not connected to an electricity source and traditionally do not have any digital devices attached to them. Information on the route, contained goods, and maintenance status used to be only available as paper printouts associated with the wagon. Only in recent years, rail companies increased efforts to digitize and automate their operations. IT systems were expanded and new technologies introduced. Equipping wagons with Global Navigation Satellite Systems (GNSS) receivers and radio-frequency identification (RFID) are typical initiatives (Šimeková et al. 2013; Balog and Mindas 2017). These technologies can make processes more efficient and provide new information, especially abroad where data availability has been low and delayed.

### 1.3 Past research and knowledge gap

The importance of shifting passenger and freight transportation from road to rail with regard to the sustainability challenges of the 21<sup>st</sup> century has been discussed widely (Sims et al. 2014; Suchanek 2017; Marinov 2018). Transportation plans on all political levels acknowledge this by aiming for modal shifts towards more rail in the future (e.g. European Commission 2011). The goal of the European Commission (EC) is to shift 50 % of current long distance road freight to rail or waterborne transport by 2050 (European Commission 2011). Steps that were defined to reach that goal include further development of rail freight corridors within Europe, better coordination between stakeholders, better intermodal integration, deregulation, and funding for research and innovation (European Commission 2011; Islam et al. 2016; Chen et al. 2018). These measures aim to make rail freight more reliable, faster, cheaper, flexible, and customer friendly.

Next to political prioritization, advances and price cuts in technology promote innovation in the rail sector. GNSS like the Global Positioning System (GPS) is one of the technologies that obtained a central role in the transportation industry. Over the past decades, the amount of created GNSS data has greatly increased and with it the need to extract meaningful information from it. Statistical analysis, machine learning, and data mining are typical big data analysis tools used to reach that goal. Studies have shown that GNSS big data analyses can filter out relevant information and create new knowledge from large datasets. Zhang and Mi (2018) analyzed GNSS data obtained from the main Shanghai bike-sharing system to evaluate the amount of carbon emissions saved by its usage. Dong et al. (2019) used GNSS trajectories of taxis in combination with passenger capacity data to extract times and locations in which the

operation efficiency could be improved. In the rail industry, research including location data has mainly focused on optimizing routing and scheduling (Goverde and Hansen 2000; Daamen et al. 2008; Medeossi et al. 2011). In those cases, the location of entire trains and not necessarily of single wagons is of interest.

Pilot studies for individual wagon tracking have been conducted since the 1990s (e.g. Welles and Hershey 1997). The results show that wagon tracking offers opportunities in different respects including internal fleet-management (Ballis and Dimitriou 2010), offering real-time product-tracking to customers (Cosulich et al. 2006), and maintenance logistics (Mirzabeiki et al. 2012). Despite these benefits, the implementation and further development of applications in the railway industry have been slow (Mirzabeiki et al. 2012).

The majority of studies concerning rail transportation using big data focuses on maintenance (Ghofrani et al. 2018). The topics of condition-based and predictive maintenance have been studied most (e.g. Sammouri et al. 2013; Li and He 2015; Papaelias et al. 2016; Yin and Zhao 2016; Zarembski et al. 2016), while corrective maintenance (Nunez et al. 2014; Corbetta et al. 2015), which is performed after a defect occurs, has been analyzed less (Ghofrani et al. 2018). GNSS data is not usually used in these studies. Instead, data from sensors on the wagon or track is evaluated. This can encompass acoustic bearing detectors, hot box detectors, or wheel impact load detectors (Ghofrani et al. 2018).

Although research using GNSS big data in the railway industry has been conducted, the focus has been on scheduling, route planning, and travel demand (Ghofrani et al. 2018). No study that uses location data to derive information on international maintenance is known to the author. Clearly, there is still potential for further research on applications of wagon GNSS location data. This study wants to contribute to this demand.

#### 1.4 Research objective

DB Cargo equipped the majority of its fleet with telematics units over the past years, thus enabling the tracking of wagons using GNSS. This data is used in this thesis to derive critical information about damaged wagons through a combination of theory-based and data-driven methods. The aim is to determine the day when a damaged wagon entered a maintenance facility abroad and when the wagon was back in operation. This task is divided into two main steps: Firstly, maintenance facilities needed to be identified and georeferenced. This was done using geofences, a “virtual geographic boundary [...] that enables software to trigger a response when a mobile device enters or leaves a particular area” (*Oxford Dictionary of English* 2010). Secondly, an algorithm that uses the maintenance facility geofences, the GNSS wagon data, and additional relevant information was developed to evaluate if and how long a wagon was in a maintenance facility. Finally, the effectiveness of the created algorithm needed to be evaluated.

Using existing GNSS data to track and localize damaged wagons is a strategic and efficient tool for a more transparent European rail industry. Information about the current damage status of a wagon is important for fleet planning, communicating delays to customers, and ensuring prompt repairs. It supports an efficient use of limited assets, cost reductions, potentially additional revenues for the wagon owner from compensation, and indirectly a more reliable and attractive mode of transportation to customers. This thesis will also contribute to geographic information systems (GIS) research by applying spatial analysis and big data tools to a new context and testing its limits with regard to automated geofence creation around existing, but insufficiently referenced geographic objects like cargo rail maintenance facilities.

To limit processing times and the amount of results to be evaluated, the analysis of this report will focus only on Sweden and Italy. These countries were chosen as most reported damages for DB Cargo wagons happen within them. The goal is however, to create a methodology that can be applied Europe-wide. The main research questions to be answered are as follows:

1. What methodology is appropriate for automatedly creating geofences around existing but insufficiently referenced geographic objects at a large scale?
2. With which accuracy can the created algorithm approximate the day of the operations release of a damaged wagon by using maintenance geofence exits?
3. Are there statistical differences in the accuracy of the approximate operations release created with the algorithm for different maintenance geofences based on their relative location (e.g. within a rail freight yard, separated from main tracks)?

It is expected that geofences generated with spatial analysis operations can detect damaged wagons correctly in certain cases. The relative location to other rail infrastructure, country-specific maintenance handling (e.g. mobile maintenance), and data quality will influence the accuracy of the result. The analysis will show for which cases the damage status can be detected reliably and for which cases additional, possibly manual efforts are necessary.

## 2. Background and related work

### 2.1 Railway industry in Europe

Economic viability is essential for every business. Profits need to be earned to keep up production and operations and invest into the future. The railway industry has a long history of financial hardship. Its beginnings, however, were very prosperous. British *laissez-faire* policies and eager investors led to immense private investments into emerging rail businesses in the 19<sup>th</sup> century in Great Britain. Although the boom declined with the burst of the rail stock market bubble in the 1840s, large private investments remained essential for rail development in other countries (Odlyzko 2010). Private investors like the Rothschild family were drivers for railroad construction across borders in Europe and beyond (Roth and Dinhobl 2008). But even in its early stages governments got involved in rail network development in some European countries as economic and military interests motivated the construction of uneconomic routes through governmental subsidies (Millward 2005).

The political interest of an extensive rail network with uneconomic routes, an increase of costs, especially in labor, the abuse of monopoly power in pricing, and negative impacts from the Great Depression are factors that first increased regulation and drove most railway companies into state-ownership until the 1930s (Millward 2005). The growing competition with road transportation and the destruction of rail infrastructure in the two world wars sealed the end of the railroad era.

Although private rail operations have been emerging since the before-mentioned EU market liberations, most of Europe's rail network is still operated by state-owned, subsidized companies. This links investments and innovation to political intent and tax revenues which constitutes one of the main reasons for the lack thereof today. The question is however, if economically profitable operations are even possible or desirable in the rail industry. The large amounts of expensive assets and personnel increase costs while low prices are needed to attract customers and compete with road transportation. Luckily for the rail industry, the climate crisis presents new incentives for governments to invest into infrastructure and innovation and potentially start a second bloom of the railway.

#### 2.1.1 Wagon telematics units

Equipping wagons with sensors is one of the most popular current digitalization initiatives in the cargo rail industry. Telematics units are mobile technical devices that get fastened to a wagon. The main unit contains a mobile communications module that enables the device to send the recorded information via the mobile phone network (Global System for Mobile Communications, GSM). The telematics system monitors different parts of the operation and infrastructure, e.g. the position and acceleration of a vehicle, environment variables like the temperature inside a wagon, or infrastructure conditions, continually and in real-time (Ramirez

et al. 2020). Typical sensors and devices included in the rail industry are GNSS receivers for localization and shock detection, temperature, or load sensors (Ramirez et al. 2020). Sensors are either included in the main unit or attached separately on the vehicle. All collected information is combined at the telematics unit and sent to the device manufacturer. After processing the raw data, it is transferred to the logistics company or other users. These systems are already widely used in the automobile and trucking industries. According to a Deloitte study by Schiller et al., every second truck was equipped with telematics units in the Triad markets (North America, Europe, Japan) in 2016. It is expected that all trucks will be equipped by 2026 (Schiller et al. 2017). Equipping European freight rail wagons with similar technology has been initiated more widely only in the past five years (DB Cargo AG 2021a).

The implementation of telematics units for freight rail is more complex than for passenger rail or road freight vehicles because cargo wagons are not connected to a power source. Solar-powered batteries are used in most systems, which highly limit the available electricity, however (DB Cargo AG 2021a). While GNSS units determine and send the position of a vehicle continuously for cars, trucks, and passenger railways, this is not feasible for freight rail wagons. Instead, the location and other status information is only sent in specified intervals. Even though the data density is decreased, devices can still encounter battery problems, e.g. if it is cloudy for a long time or if the wagon is stationed under a bridge or other sun-shielding infrastructure (DB Cargo AG 2021a). This can lead to missing data and decrease the data quality.

Most telematics units use several available GNSS, e.g. GPS, Galileo, GLONASS, or BeiDou. The horizontal accuracy of most GNSS receivers is approximately 5 m in good conditions (European Space Agency 2011a, 2011b; U.S. Space Force 2021). It can deteriorate if surrounding buildings, trees, or water interfere with the signal (Merry and Bettinger 2019). If no GNSS signal can be obtained, for example because not enough satellites are reachable, the location of the nearest cell tower used to transmit the data can be used to approximate the actual location (DB Cargo AG 2021a). The accuracy of this data is much lower than with GNSS, however.

### 2.1.2 Wagon telematics at Deutsche Bahn Cargo AG

DB Cargo AG is the largest national cargo RC in Europe (Railway Technology 2018). It operates in 17 European countries and manages over 90,000 wagons (DB Cargo AG 2021b). The goal of the Wagon Intelligence project of DB Cargo is to transform their fleet into “intelligent freight wagons” by equipping each wagon with sensors. The main component is a telematics unit that measures the GNSS location, detects movement and shocks. A solar panel enables its usage over several years. Other passive components that can record additional information include near-field communication (NFC), ultra-high frequency (UHF) tags and sensors for humidity, temperature, dew point, and load-detection. The main unit sends the data in specified intervals or at special events (e.g. the wagon stops) using the mobile telecommunication network. During pre-processing, the telematics data is connected to the

according wagon and enriched with existing business and wagon data from internal systems. Subsequently, it can be used for data analysis or the creation of internal or client-targeted tools. Examples for these applications are live-tracking of goods for clients, automated maintenance scheduling, demurrage of wagons at the client's property, or condition-based maintenance.

## 2.2 Big data and data-driven research

*“Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.”* – Chris Anderson, 2008

A wide discussion was started in the scientific community in 2008 with an article by then *Wired* magazine editor-in-chief Chris Anderson. He predicted the end of science as we know it: theory-based and hypotheses-driven. Collecting, storing, and evaluating data has never been as easy as today. Petabytes of information are at hand, waiting to be analyzed. But does this entail the end of hypotheses testing in research? Data-driven science uses algorithms and statistical tools to derive knowledge from massive data amounts (Kitchin 2014). It starts with data, not with theories. It therefore follows inductive reasoning while theory-based science follows deductive scientific reasoning. Miller and Goodchild (2015) describe this new research paradigm very pointedly as an inversion of the “classic sampling problem where we identify a question and collect data to answer that question. Instead, we collect the data and determine what questions we can answer”. Characteristics of data-driven research include an explorative approach opposed to a predictive one, analyzing populations instead of samples, handling unstructured, “messy” data, and focusing on patterns and correlations instead of causalities and universal truths (Shmueli and Koppius 2011; Strasser 2012; Jagadish 2015; Mazzocchi 2015; Miller and Goodchild 2015).

This new approach makes it possible to discover and answer questions that we would otherwise not even think to ask (Mazzocchi 2015). Analyzing data on entire populations instead of supposedly representative samples gives a more inclusive picture of reality. Instead of controlling and simplifying conditions, data-driven research embraces the complexity and “messiness” of data as a more accurate depiction of nature (Miller and Goodchild 2015). It does not require finding universal truths but fosters an indefinite path to increased knowledge as new inferences can alter inferences already made (Malle 2013).

Big data plays an essential part in data-driven science. It is characterized by the five V's: volume, velocity, variety, veracity, and value (Jin et al. 2015). Next to the large data amounts (volume), big data consists of many different, possibly unstructured data types (variety). It is created continually and often requires real or near real-time processing (velocity). The large data amounts lead to uncertainties and inaccuracies within the data (veracity), but they also provide an enormous potential to new knowledge and insights (value) (Ishwarappa and

Anuradha 2015). Sources for big datasets can be social media postings, telecommunication and transportation data, search engine queries, online shopping statistics and many more.

These new, ever-growing, and complex data volumes create new challenges. An adequate IT infrastructure is necessary to process the data. Analysis tools need to be modified and combined to assess diverse data types (Jin et al. 2015). Researchers and professionals need to keep up with the development of new tools and technologies to properly handle and evaluate the data. Data storage and cataloging methods need to be updated to manage unstructured data like video, audio, and text files. Lastly, data security and personal data protection laws need to be upheld (Gaur 2020).

Geographers have handled large data amounts for half a century. Long before the age of GNSS modules in mobile phones and geo-tagged social media data, researchers had to cope with big data in remote sensing. Since the start of the Landsat program in the 1970s evaluation and storage methods developed together with the increasing data volume (NASA 2021). Other geo-referenced data sources include GNSS trackers in vehicles, mobile phones, and infrastructure, RFID tags and social media posts (Miller 2007; Miller 2010; Sui and Goodchild 2011).

The discussion on the possibly new paradigm of data-driven science is still ongoing. Mazzocchi (2015) predicts that new knowledge will be inferred from automated data mining. But he also states that the human creativity cannot be replaced by statistical models and big data analyses, and he underlines the advantages of iterative cycle of inductive and deductive phases. Strasser (2012) emphasizes that inductive reasoning has been a part of scientific practice for centuries. Kepler's laws on planetary motion for example were based on observational data. Data-driven science differentiates itself by cross-disciplinary data evaluation, the usage of advanced statistical tools, and diverse data sources, however (Strasser 2012). Many researchers and experts stress the importance for data-driven research to produce and enforce standards and contribute to the development of methodological frameworks (Strasser 2012; Das et al. 2015; Maass et al. 2018).

This thesis combines the hypothesis-driven and the data-driven approach. The data is the driver of the research asking the question *What knowledge can we deduce from cargo wagon GNSS data?* The hypothesis is that this data will be useful to provide additional information on the damage state of cargo wagons.

## 2.3 Geographical Information Systems

### 2.3.1 Geofencing

A GIS is “a computer system that analyzes and displays geographically referenced information. It uses data that is attached to a unique location” (USGS 2021). GI systems offer a great variety of data processing, analysis, and visualization tools. Geofencing is one of those tools. GNSS inaccuracies make it practicable to use an area instead of a point to determine when an object is at a certain location. Geofencing has been used in many application from notifications if a

shipment reaches a point of interest in logistics (Reclus and Drouard 2009) to tracking fans that orientate their air flow towards a person entering a certain area (Liu et al. 2017) to alleviating gambling disorders by app-based notification when a patient gets too close to a casino (Coral et al. 2020). Whenever the arrival, departure or stay of mobile objects like vehicles, persons or containers at certain locations are of interest and GNSS information is available, geofences are the appropriate tool. This is the reason why they are used in this study. Geofences enable to determine when a wagon arrives, stays, or leaves a cargo wagon maintenance facility.

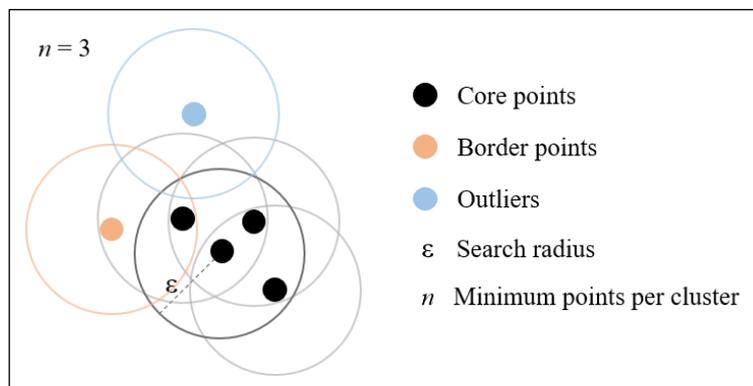
In most applications, the areas of interest are tagged manually by drawing borders around them, for example in web GIS interfaces, or by determining the relevant coordinates for the fence in another way (Cardone et al. 2014; Coral et al. 2020). In other cases, geofences are created automatically around new points of interest, e.g. while determining the most ideal parking spots for shared bikes or electric scooters in a city with location-allocation models (Cheng et al. 2019; Pérez-Fernández and García-Palomares 2021). But what if existing buildings or areas should be referenced automatically? Large crowd-sourced databases like OpenStreetMap (OSM) provide detailed information on infrastructure. As infrastructure elements are organized by assigning appropriate tags, similarly-tagged items can be extracted easily. If geofences around every school in Germany are needed, a query with the tag *school* should give a comprehensive subset of all relevant buildings. The result is only as good as the quality of the tags, however. If an OSM contributor forgot to assign the tag to a school, it will not be included in the result. The completeness and accuracy of the OSM data varies throughout the world. In Europe, the OSM community is very active. In other parts of the world like Asia and Africa, less information has been digitized (Barrington-Leigh and Millard-Ball 2019). Additionally, less widely available information is less likely to be included in these databases, independent on the location on Earth. While evaluating the data quality of tags of cargo rail maintenance facilities in OSM, the tags of known maintenance buildings were compared. The value of the building tag ranged from *railway* to *industrial* to *yes* only indicating that it is a building. The knowledge about the exact use of a building is often limited to the people working there. If those people do not contribute that knowledge to the database, it cannot be found therein. This thesis develops and tests a methodology for creating geofences around existing, but insufficiently referenced geographic objects at larger scales where manual geofencing would be too cumbersome and knowledge about the objects is not centralized. No study that implemented such an approach could be found in the literature.

### 2.3.2 Clustering

Another GIS tool used in this thesis is clustering. Clustering is not limited to geographical applications, it constitutes the process of grouping similar objects (Kresse and Danko 2012). Similarity can be present in attributes of the data objects or in location. Similar objects always lie close to one another either in actual space (location similarity) or in multivariate space

(attribute similarity). The nearest neighbor index or Moran's I can be used to determine if a dataset is clustered or spatially random (Chang 2016). If a dataset is clustered, many different approaches can be used to identify the clusters. They can be grouped into hierarchical and non-hierarchical clustering methods (Ma and Chow 2004). Hierarchical methods start with  $n$  clusters for  $n$  data points. Step by step similar clusters are merged until no similar clusters remain (Ma and Chow 2004). Data points that have been grouped into one cluster will remain in the same cluster. Examples for these kind of clustering algorithms are BIRCH (Zhang et al. 1996) and CURE (Guha et al. 1998). Disadvantages of this approach are the high computational effort required and the resulting long computation time for larger datasets.

Partitioning type clustering is one type of non-hierarchical clustering methods (Ma and Chow 2004). It starts with a user-defined number of cluster seed points. Data points are assigned to the nearest cluster point. The location of the cluster seeds is changed iteratively until the groups are stable and no improvement of a chosen metric, e.g. variance, can be obtained. An example for this type of clustering approach is k-mean clustering (Hartigan 1975). An advantage of k-means and other partitioning clustering methods is their linear complexity and the resulting short computation time (Maimon and Rokach 2005). However, they do not function well with arbitrary cluster shapes (Ma and Chow 2004). Another disadvantage can be that the number of clusters needs to be known in advance. Some algorithms approximate the optimal number of clusters, e.g. Ishioka (2006).



*Figure 1: Visualization of functionality of DBSCAN clustering. The circles indicate the search radius  $\epsilon$ . The point with the darker circle is the start point. As more than three other points are within the  $\epsilon$  of the start point, a cluster is created. The border point lies within  $\epsilon$  of a center point but does not have  $n$  points within  $\epsilon$ . It is part of the cluster but new points within  $\epsilon$  of the boundary point are not added to the cluster. The outlier is more than  $\epsilon$  away from any other point.*

Density-based methods are another example of non-hierarchical clustering approaches. They analyze the density of the data points to determine clusters (Ma and Chow 2004). Density-based spatial clustering of applications with noise (DBSCAN) is one popular density-based clustering method. It starts at a random data point and selects all data points within a specific search radius  $\epsilon$  around that point (Figure 1). If the number of points within  $\epsilon$  is above a certain threshold  $n$ , a cluster is created. All points of that cluster that have at least  $n$  points within distance  $\epsilon$  are core

points. Points that are within distance  $\varepsilon$  of a core point but do not have the minimum number of points close by are border points. For every core point in the cluster the surrounding points within distance  $\varepsilon$  are added to the cluster. Once no more close points can be found, a new point, not yet in a cluster is chosen. Points that have a distance greater  $\varepsilon$  to any cluster point are classified as outliers.

The minimum amount of points per cluster  $n$  and the search radius  $\varepsilon$  need to be defined by the user. Campello et al. (2013) developed a hierarchical DBSCAN (HDBSCAN) which only needs a minimum cluster size as input. Density-based clustering methods work well with arbitrary cluster shapes (Ma and Chow 2004). DBSCAN's ability to detect outliers is a very useful feature. Its performance decreases if the density varies only little throughout the dataset or if the density of the clusters vary highly. In these cases, it is difficult to set the  $\varepsilon$  value.

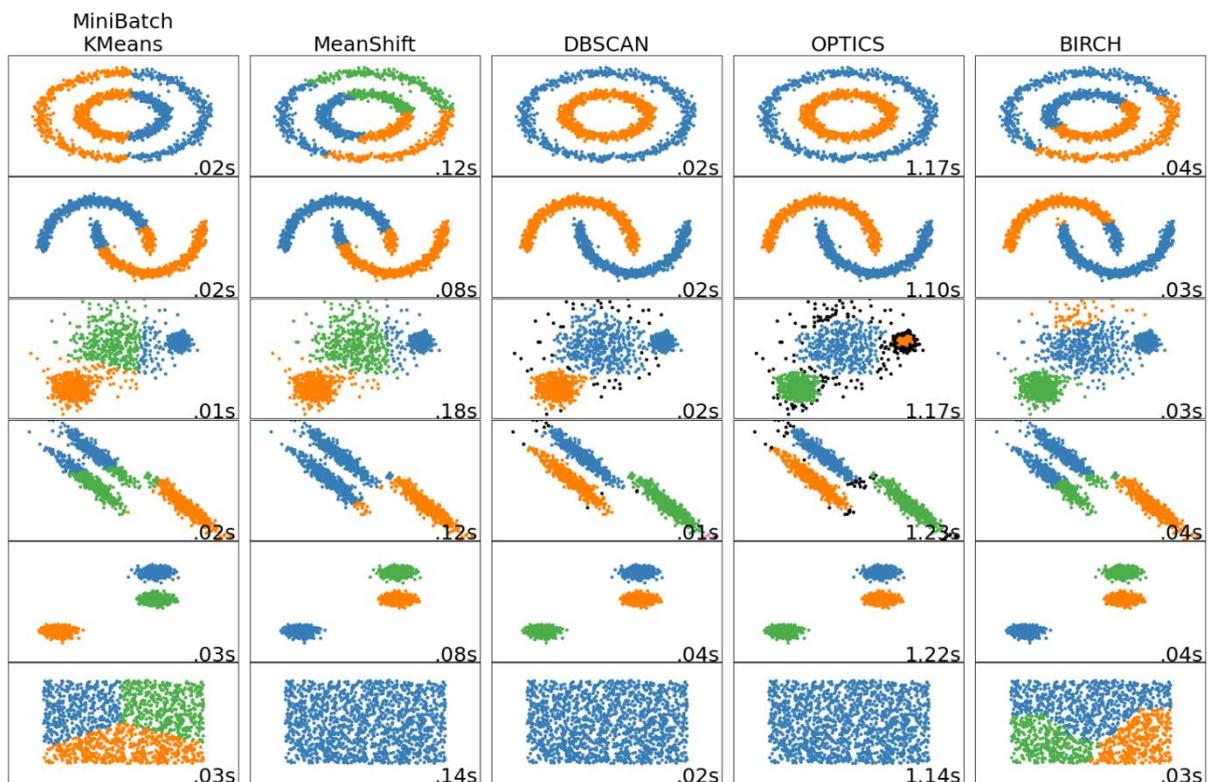


Figure 2: Behavior of different point clustering algorithms for point dataset examples and their computation time created with Scikit (Scikit-Learn Python Library 2021).

Gülagiz and Sahin (2017) compared different clustering methods and found that k-means clustering is the fastest approach and DBSCAN the most accurate. However, every clustering approach has strengths and weaknesses. Figure 2 shows how different clustering algorithms behave with different kinds of datasets. In the end, the most appropriate clustering method for the data at hand needs to be chosen.

### 3. Methodology

This thesis consists of two main parts. To reach the aim of creating an algorithm for the detection of damaged wagons in maintenance facilities, these maintenance facilities needed to be identified and georeferenced first. Identifying existing geographic objects can be performed in different ways. Either data on the objects is already accumulated in some form, e.g. as a list, a database, or expert knowledge, or it must be acquired first, e.g. by using GIS techniques like remote sensing or spatial analysis. For this study existing data in the form of unstructured lists, database entries, and structured vector data was used to create the maintenance facility geofences. Knowledge from workers in maintenance facilities was used to assess the quality of the geofences. The effectiveness of analyzing aerial or satellite imagery was deemed too low for the detection of rail maintenance facilities. Reasons include the indistinctiveness of rail maintenance facilities compared to other industrial buildings or rail storage facilities and known difficulties related to applying raster analyses over large areas like Europe (e.g. cloud cover, availability of imagery from the same source, time, and in the same quality to compare spectral bands).

After the geofences were created and validated, a Python algorithm that evaluates if and how long a wagon was in a maintenance facility geofence was created. The quality of the algorithm was assessed by comparing the day a wagon left a relevant maintenance geofence that the algorithm computed to the day of the actual operations release (OR) which is sent by an RC after the repair is concluded. Details on the data and methodology are described in the following sections.

#### 3.1 Input data

Data from four main sources were used in the analysis:

- Wagon telematics data collected by the telematics units attached to each wagon
- Information about GCU damages stored in the WADIS database (Wagon Disposition and Information System)
- Unstructured lists with addresses from maintenance facilities in Europe obtained from DB Cargo and the Association of Freight Wagon Holders
- Openly available infrastructure data from OSM

While this study only examines Sweden and Italy, the created methodology should be applicable to all European countries. This was considered during the data acquisition and preprocessing steps. All geographic data was processed with the European projection ETRS89 LAEA (EPSG:3035) which fits data covering most of Europe best.

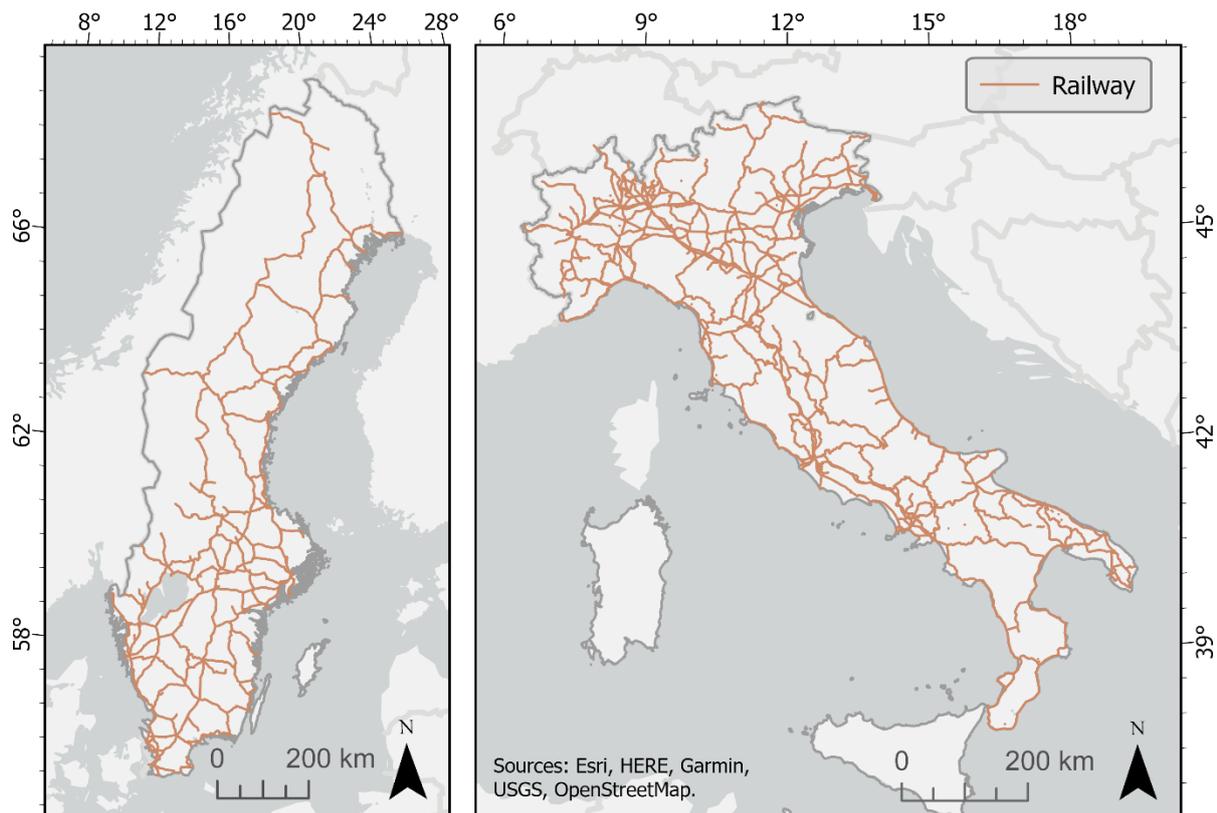


Figure 3: Study areas with national rail network. Left: Sweden with SWEREF99 TM projection. Right: Italy with RDN2008 projection.

The main rail network of the two study areas Sweden and Italy is shown in Figure 3. Sweden's rail network consists of few main rails connecting the most important cities and industrial areas. The principle cargo RC is Green Cargo AB. There are only few railway maintenance providers that operate most maintenance facilities, SweMaint and EuroMaint being two examples. The rail connection to Germany through Denmark and shipment options by ferry, for example from and to Trelleborg, are some of the most important international routes. Sweden and Italy are connected through the European Scandinavian-Mediterranean rail freight corridor (European Commission 2021). Italy is also part of the Baltic-Adriatic corridor running through Austria and Slovakia to Poland, the Rhine-Alpine corridor to the Rotterdam, and the Mediterranean corridor from Spain to Hungary. The Italian rail network is very dense in the North but less so in the South. Several cargo RCs operate in Italy, the main ones being DB Cargo Italia and Rail Traction Company. Maintenance operations are less centralized in Italy with many smaller providers, some examples being Italy Rail, Cosmef, and ECMS.

### 3.1.1 Wagon telematics data

As described in more detail in section 2.1, the telematics units mounted to each wagon send information about the wagon in certain intervals. Different wagon types are equipped with different sensors and modules. Every wagon with a telematics unit has at least a GNSS module and an acceleration sensor. Additional instruments include temperature and load-detection

sensors. The main telematics unit sends the recorded information every 10 minutes while the wagon is moving, every 24 hours while it is parking, or at events, e.g. the movement state changes.

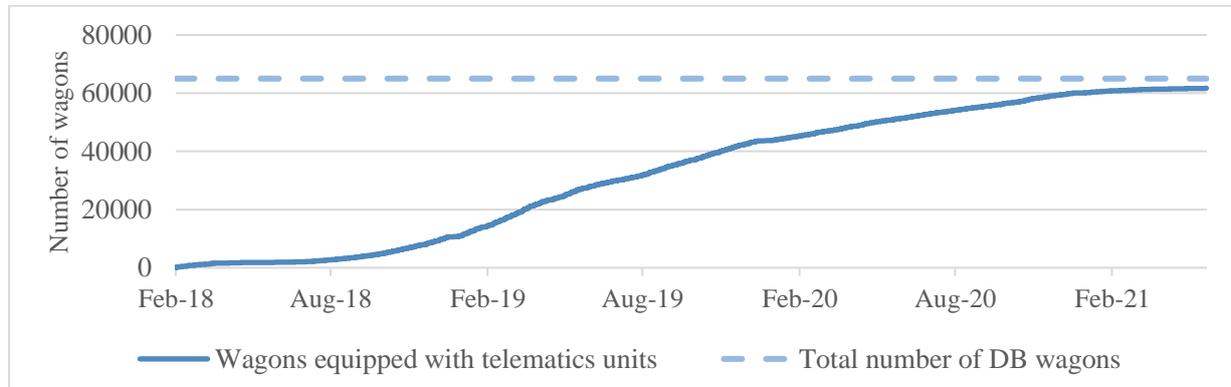


Figure 4: Number of wagons owned by DB Cargo that are equipped with telematics units 2018 to 2021.

The data is stored in a cloud environment using Hadoop Distributed File System (HDFS). Data analytics can be performed directly in the cloud environment with e.g. Python. All telematics data collected since the start of the project in 2018 is available. As wagons were continually equipped with telematics units, the number of wagons sending data and the overall daily data amount increases throughout time. Since the beginning of 2021 most wagons have been equipped with telematics units (Figure 4). The number of data points fed into the database every hour usually ranges between 70.000 and 140.000, amounting to approximately 2.5 million data points a day. As this value highly depends on the number of wagons that are moving, it fluctuates considerably. The relevant information available on each data point is portrayed in Table 1.

Table 1: Relevant information provided by the telematics dataset for each data point.

Information	Data type	Details
Wagon number	Integer	---
Wagon type	String	According to International Union of Railway (UIC) code
Closest railway station	String	---
ID of railway station	Integer	---
Distance to railway station	Float	In kilometers
Current order details	Integer	E.g. order ID, type of goods
Latitude	Float	In decimal degrees
Longitude	Float	In decimal degrees
Localization source	Integer	GNSS or GSM
Timestamp measure	Date/time	Time, measurements were taken
Total mileage	Float	Total distance travelled
Mileage delta	Float	Distance travelled since last transmission

The *closest railway station* field indicates the railway station that is closest to a wagon. To give a reference to the wagon coordinates, DB Cargo created a Voronoi diagram from known railway stations in Europe. For each railway station, a point coordinate is used as a center point of the

Voronoi diagram. This enables to determine the closest railway station to any point in Europe. As the area of each railway station cell can vary greatly depending on the density of stations, the distance to the station point is given as well in the field *distance to railway station*.

### 3.1.2 Damage information in WADIS

If a damage occurs abroad, a damage protocol needs to be sent to the wagon owner as stated in the GCU. Currently, they are usually sent as PDF or text documents via email. At DB Cargo, these are semi-automatically collected in the WADIS database. A program extracts information like the wagon number, the time and place of damage, and damage code automatically from the PDF. A staff member then checks the result and adds data that could not be automatically read.

All damage protocols that were received by DB Cargo since WADIS was created in July 2016 are stored in the database. They amount to over 85,000 damage protocols. The data can be searched with an in-built filter and the results can be downloaded. The relevant information included in the download is shown in Table 2.

Table 2: Relevant information included in downloads from WADIS.

Information	Data type	Details
Wagon number	Integer	---
Company name	String	RC that sent damage protocol
Damage protocol ID	Integer	---
Day of damage	Date	---
City of damage	String	Not always same syntax for the same location
Country of damage	String	Country code (e.g. SE for Sweden), only included for algorithm development, not available during geofence creation
Damage descriptions	String	Same text syntax for each damage code
Form K	Binary (0/1)	If 1, repair in Germany
Form M	Binary (0/1)	If 1, only minimal damage, no immediate repair
Damage cause	Integer	Wear / abuse / third party / cause not ascertainable
Cost estimate	Binary (0/1)	If 1, cost estimate was received
Spare parts	Binary (0/1)	If 1, spare parts were needed for repair
Operations release (OR)	Binary (0/1)	If 1, operations release was received
Date of operations release	Date	---

The quality of the data in WADIS differs. Open text fields are naturally prone to syntax errors. Thus, the field for the city where the damage was detected is not always comparable or accurate. Typing errors in the date of the operations release (OR) can also lead to cases in which the date of the OR is before the damage day. There seem to be internal database errors as well, however. There are for example cases for which the field *Operations release* indicates that there is an OR, although there is no date given. Other times, the date of the OR contains logical errors, e.g. due to a database error which returns 01.01.1970 as the OR date. While working with the data, these quality issues need to be considered.

### 3.1.3 Addresses of maintenance facilities in Europe

Reliable and complete data on cargo rail maintenance facilities in Europe is the base for creating accurate geofences. As the names of these facilities differ widely throughout Europe, they cannot be found simply by doing a Google Maps search. Additional data, e.g. in form of a list with names and addresses of maintenance facilities, is necessary. It was difficult attaining this data, however. DB internal lists were incomplete and often without street addresses for international maintenance facilities. Other sources like the GCU signatories contained no information about which business performs maintenance operations. Even if an address was provided, it was often for the headquarters, rather than the maintenance facility. The use of open-source data like OSM was considered but as the tags of rail maintenance facilities are very inconsistent throughout Europe, no meaningful information could be extracted through this means.

The Association of Freight Wagon Holders in Germany (Verband der Güterwagenhalter in Deutschland e.V., VPI) combines over 250 companies in the railway industry not only in Germany but in Europe. They work together on standards and data exchange and operate 199 maintenance facilities in Europe. A list of these facilities including street addresses was obtained. This is the main source for maintenance facility addresses. As DB Cargo is not a member of VPI, its maintenance facilities are not included in the list. Thus, it was decided to use both the VPI and the DB data to find most of the maintenance facilities in Europe. 199 addresses from VPI and 163 from DB Cargo were used as inputs. The different quality and information contained in the two datasets led to the application of different analysis steps for each of them.

### 3.1.4 Infrastructure data from OSM

To perform spatial analyses, infrastructure data for Sweden and Italy was needed. This encompasses firstly the rail network and secondly buildings data used to later extract possible maintenance facility buildings. The OSM data was explored using its web-based data mining tool Overpass turbo and the OSM website. As the tags used by OSM to classify rail industry buildings are not uniform throughout Europe, it was not possible to merely download rail-related buildings. Instead, all buildings were downloaded and filtered during preprocessing (see section 3.2). As the datasets containing all buildings within a country were too large to extract through Overpass, they were obtained through geofabrik.de, a website that provides current OSM data downloads.

The downloads were performed for Italy and Sweden. As the data quality of OSM is very high throughout Europe and the proposed methodology should be applicable for all of Europe, OSM is the most appropriate source for this project.

### 3.2 Data exploration and preprocessing

The goal of the data exploration phase was to gain insights into the data used for the analysis. As an explorative research approach was used, it is important to know what information is available, in which formats and in what quality. An examination of the WADIS data gave interesting insights into damage cases in Europe and helped to define the scope of the project.

The 199 maintenance facilities from VPI and the 163 from the DB Cargo datasets needed to be aligned to work efficiently with them later. After adjusting the data formats and structure, the datasets were combined and duplicates were removed. The number of entries per country in the VPI, DB Cargo, and the combined datasets can be seen in Appendix A. The final dataset contained 314 maintenance facilities in Europe with 12 in Sweden and 16 in Italy.

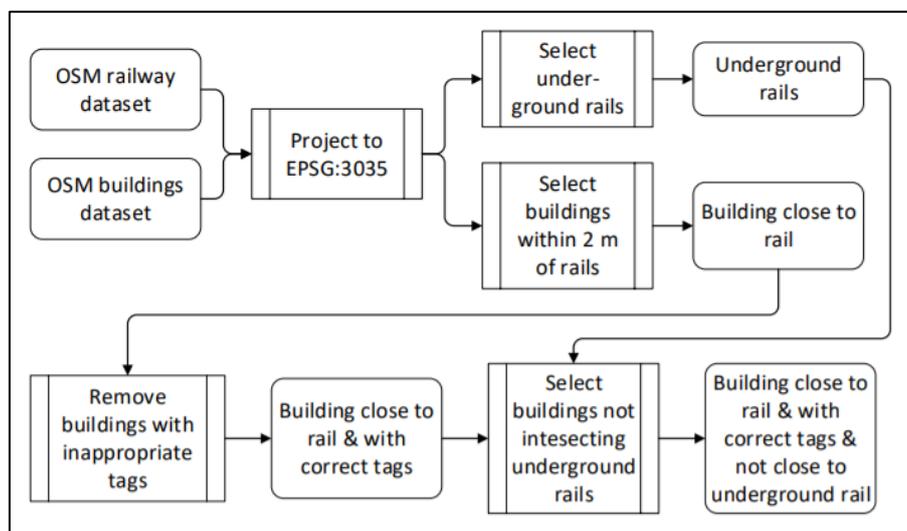


Figure 5: Steps performed to preprocess OSM buildings data and filter out relevant buildings.

The OSM buildings data also had to be preprocessed. As described in section 3.1.4, it was not possible to filter the buildings data effectively with Overpass. Consequently, all buildings for the selected area (here described exemplary for Sweden) were downloaded. For Sweden, this download included approximately 2.4 million buildings. The goal was to extract buildings that are possibly being used for railway maintenance. Rails connected to maintenance facilities usually go into the maintenance building. Satellite imagery is commonly used to digitize the data available at OSM. Image resolution deficits and inaccuracies of OSM mappers can result in rails disconnected to buildings in OSM. Thus, buildings within 2 m of rails were extracted. This limited the number of buildings to about 1,000. Then, the tags of the buildings were checked. Buildings containing inappropriate *use* tags like *kindergarten* or *apartment* were excluded. Further, buildings that are close to underground rails were removed using the rail tag *tunnel*. This left only 650 buildings in Sweden that were used for further analysis. The steps performed are visualized in Figure 5. For Italy, this process resulted in approximately 3,000 relevant buildings.

As maintenance facilities are located at minor rails and not on the main tracks, the railway data was filtered to not include the *usage* tags *main* and *branch* before the relevant buildings data was selected. This decreased the number of included rails by approximately 50 % for each Italy and Sweden.

These preprocessing steps filtered out irrelevant buildings. As buildings close to address points were extracted during the geofence creation later, it was important to limit the amount of included buildings to only those, that could possibly be used as maintenance facilities.

### 3.3 Creation of maintenance geofences

As no geofences for cargo rail maintenance facilities in Europe exist openly available today, it was necessary to create them. Given the large number of maintenance facilities, it was desirable to automate this process. Drawing the geofences manually would not have been easy even if there were less maintenance facilities, however. As it was already difficult to acquire addresses of maintenance facilities, finding the actual location would require contacting many different experts all over Europe. Another motivation for the automation was to test the suitability of spatial analysis tools for this purpose.

The geofences were constructed in three main steps:

- First, geofences were created by geocoding addresses of maintenance facilities and using infrastructure data to select relevant buildings and rails near the address points.
- In a second step, another set of geofences for maintenance facilities was created using GNSS wagon data and damage information. GNSS points of damaged wagons were selected and grouped using clustering algorithms to extract areas where many damaged wagons were located.
- The last step comprised joining the two sets of geofences to improve the overall result. This approach was chosen, as no complete list of maintenance facilities in Europe could be acquired. The GNSS geofences should try to detect maintenance facilities that were not in the list.

#### 3.3.1 Geofences from address and infrastructure data

The first step in creating geofences from addresses and infrastructure data was to convert the addresses of maintenance facilities to coordinates. This operation was performed using the Google Maps geocoding function. As experience has shown that the search algorithm of Google is better than the one of OSM, a Google API was used. This search provided longitude and latitude results for all VPI facilities. Two DB entries did not get a result.

To check the quality of the coordinates, they were overlaid with a shapefile containing the country borders of the world. This examination showed that six points did not have a matching country. A manual inspection showed that errors in the original address data were the cause for

this. In three of the six cases manual searches returned the correct address. The remaining three were excluded from the analysis changing the number of address points to 311.

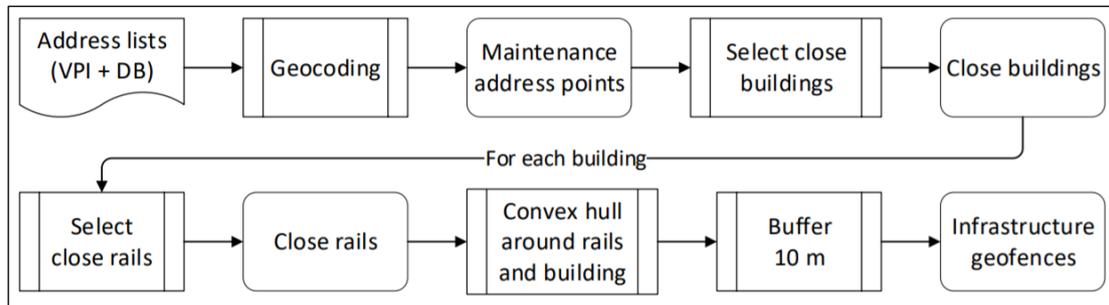


Figure 6: General process used to create infrastructure geofences from maintenance facility addresses and OSM data.

The next step was to perform spatial analysis operations to extract geofences from the address points. To make this step repeatable, the ArcGIS Pro model builder was used (Appendix B). This allows to string together different analysis tools and perform them all at once. In a first step a search radius around each address point was created. Different distances were tested and visually assessed. A radius of 500 m included the relevant rail and building infrastructure most times. Buildings that were within this distance were then selected. For some address points, especially if no street address was given, this distance did not include any buildings, however. Thus, a second search radius of 2 km was applied if no building was returned in the first examination. In Sweden 34 buildings were selected through this technique, in Italy 86.

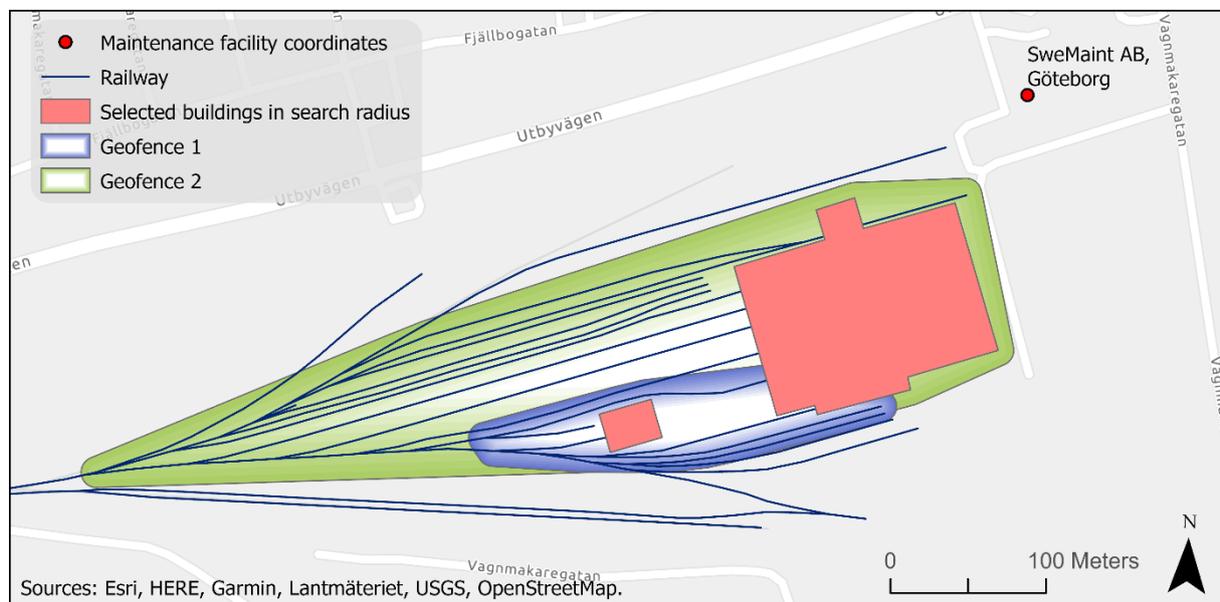


Figure 7: Example for process of infrastructure geofence computation for SweMaint AB in Göteborg. For each relevant building identified within a search distance of an address point (Figure 5 and Figure 6), an infrastructure geofence was created by selecting relevant rails and computing a bounding geometry. Projection: SWEREF99 TM.

Using an iterative process, the following operations were performed for each building in the selection. Rail features that are within 5 m of the building were selected. This distance was

chosen as relevant rails are close to the maintenance buildings as stated before. A larger distance was chosen in this instance (during the building selection 2 m were used) as some maintenance rails used to park wagons do not necessarily go into the maintenance building. Then, the convex hull for the selected rails was computed to combine them in a minimum boundary area. This polygon was combined with the building polygon. Finally, a 10 m buffer was added to the infrastructure geofence (IGF) to account for GNSS inaccuracies. Figure 6 visualizes the performed steps. Figure 7 shows the IGFs created in Göteborg at a maintenance facility of SweMaint AB.

This provides a rudimental first result of geofences for the address points. It does not, however, include any actual wagon data. Therefore, in a second step, geofences were created using GNSSs wagon data.

### 3.3.2 Geofences from GNSS wagon data

The general approach to creating geofences from wagon GNSS data was to select relevant GNSS data points and perform clustering operations on them. If only data points for wagons that are damaged are selected, they must show the location of the maintenance facilities. Clustering is an adequate technique to detect areas of interest from point data. Selecting relevant data was crucial for this process as clusters at transshipment points or client sites where wagons stay for a long time but are not necessarily damaged needed to be excluded as best as possible.

The large amounts of GNSS data requires adequate IT infrastructure and scripting techniques. Apache Spark analytics engine enables processing big data in distributed storage systems very efficiently. Its Python API, PySpark, was used to access the wagon telematics data on the HDFS and manipulate it. As only damaged wagons were of interest, the GNSS data was matched to WADIS damage data using the wagon number and time stamps. It was decided to use WADIS data from 2020 for the geofence creation, so the data from 2021 could be used for testing later. Accordingly, the GNSS data was filtered to only contain data points for which a damage entry was recorded in WADIS in 2020 and which lie in the period between the damage day and the day of the OR for that damage case. In 2020, more than 20200 damage protocols were entered into WADIS. Over 6000 of the damages occurred in Italy and over 2700 in Sweden. As the day of the operations release was important for selecting relevant data, only WADIS entries with a valid OR date were included (~ 290 for Italy and ~ 330 for Sweden). After these operations, a total of 204 000 telematics data points remained.

Clustering attempts showed that many irrelevant data points were still included in the point selection. In Malmö for example, data points could be found with damage locations in Sopron/Hungary, Hallsberg, and Norrköping (the latter two are located in Sweden but more than 400 km away from Malmö). Manual single case analyses showed that the day of the WADIS OR was not correct in almost all cases. In a sample of 13 cases, the OR date was off by +7 days on average. The cause of this are faulty manual data entries in which clerks used the date the OR was received instead of the actual OR date. If the date of the OR in the database is

later than the actual OR date, additional GNSS data points are included. Data of damaged wagons gets mixed with data of already repaired wagons which decreases the quality of the clustering. Thus, further GNSS filtering techniques were necessary to improve the result.

Damages are seldomly detected in the middle of a track. As wagons are checked for damages when they are assembled to a train, consigned to a different RC or to a client, damages are usually detected in larger freight yards or at client sites. If the damage is not minor, the wagon has to be repaired close to the location the damage was detected. With this assumption, it is possible to filter the GNSS points further, so that only points at the damage location are included. There are two cases however, that must be considered. For minor defects or more complex repairs it is more sensible to let a wagon return empty to Germany and be repaired there. In those scenarios the wagons are not repaired at the damage location and thus do not follow the assumption made above. Consequently, those cases were excluded from the analysis. There is also the case of mobile repairs, in which mobile teams repair wagons at different locations. As it is unknown at DB Cargo which maintenance companies perform mobile repairs and where exactly they operate, these cases could not be handled and will have to be considered in the evaluation of the resulting geofences.

To check if a telematics point is at a damage location, two different approaches can be used with the available information. Either the name of the damage location from WADIS and the name of the closest train station in the telematics data are compared or the WADIS damage location is georeferenced and compared to the GNSS coordinate of the wagon. As the WADIS data did not include a country code at this point, it was difficult to receive good geocoding results. It was thus decided to perform a string comparison between the damage location and the closest railway station. Both text fields were homogenized to make them comparable. The damage location string was split into single words. Some damage locations include specifications, e.g. Malmö Godsbangård. As the train station name does not necessarily include the specification, better results can be achieved by evaluating each word singularly. If one of the words had less than three characters, it was not considered in the comparison. If a damage location string was contained in the text string for the closest train station, the two were considered similar. If the exact word was not contained, an additional string comparison was performed using the SequenceMatcher from the DiffLib Python library. It is based on the Ratcliff-Obershelp pattern recognition algorithm (Ratcliff and Metzner 1988). The DiffLib documentation states that similarity values above 0.6 are usually sufficient to call two strings similar (Python Software Foundation 2021). These operations improved the result although not all inappropriate points could be removed.

At some locations a high number of damages occurred, e.g. over 100 in Chiasso, almost 50 in Malmö. As smaller clusters get more difficult to detect in very dense data, a maximum of 15 damage cases were randomly chosen for each damage location. These operations left approximately 90.000 data points for Sweden and Italy at 55 different damage locations (32 in

Italy, 23 in Sweden). In total 139 damage cases were included in the clustering analysis for Italy, and 95 for Sweden.

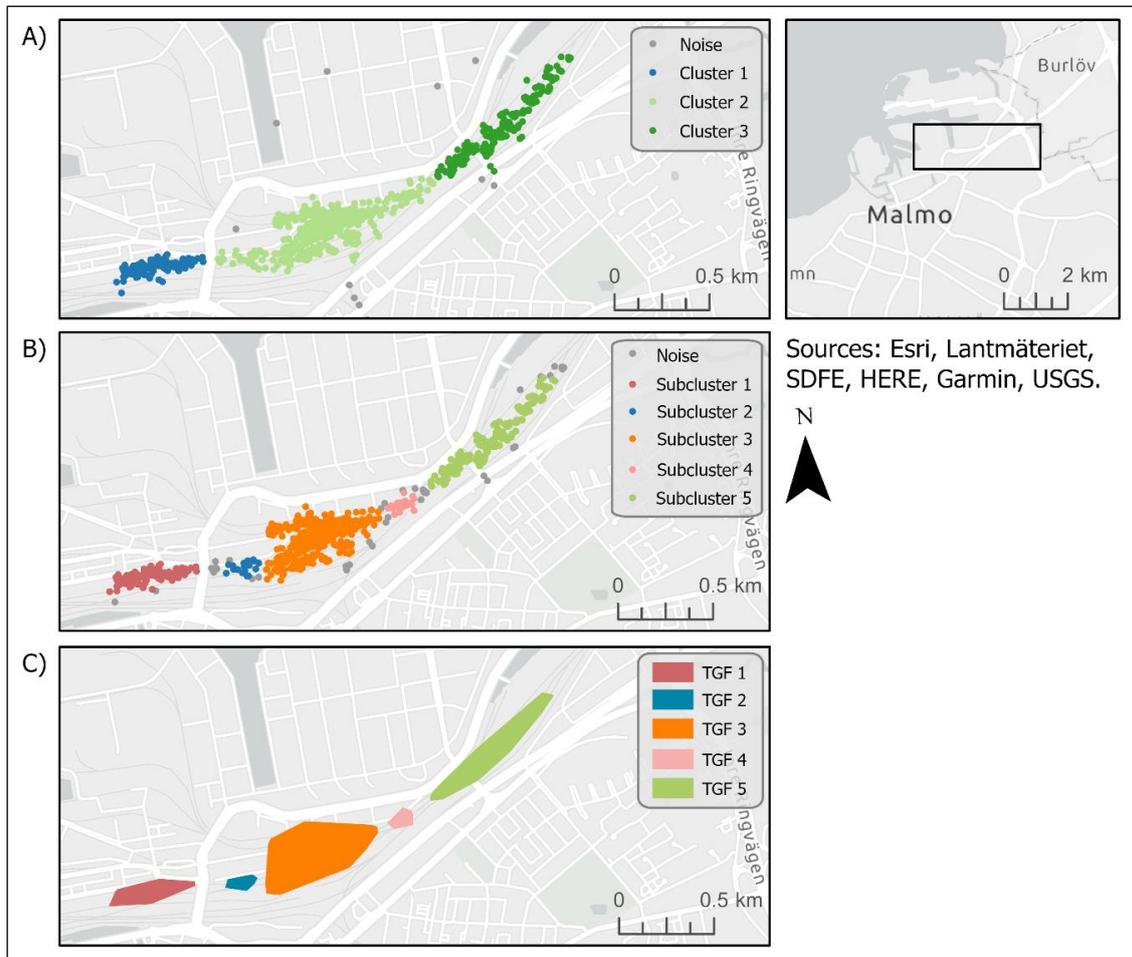


Figure 8: Process of TGF creation. A) First DBSCAN clustering with a search distance of 100 m and at least 20 point per cluster. B) Second DBSCAN clustering with a search distance of 50 m and at least 10 points per cluster, noise of first clustering process excluded. C) TGFs created from clusters using a bounding geometry. Projection: ETRS 1989 LAEA.

Different clustering approaches were considered. As the total number of maintenance facilities is unknown and the aim of the GNSS-derived geofences is to detect maintenance facilities not included in the address list, non-hierarchical clustering methods for which the number of clusters needs to be specified upfront are not appropriate. K-means clustering algorithms which approximate the optimal number of clusters were tested but the quality of the results was not sufficient. Instead, a density-based clustering algorithm was chosen. DBSCAN has many advantages that are useful for the telematics dataset. Firstly, no cluster amount needs to be specified. Secondly, it detects outliers and groups them into their own outlier class. This is very useful, as GNSS data is known to be error-prone under certain conditions. And lastly, DBSCAN is good in handling differently sized and shaped clusters. As a minimum number of points and a search radius is needed to perform DBSCAN clustering, different values were tested. In a first clustering analysis, general clusters were identified, and outliers eliminated. A minimum of 20 points per cluster and a search distance of 100 m grouped close data points adequately. This process resulted in 108 clusters in Italy and Sweden including one cluster for outliers. If no

satellite signal can be obtained, GSM is used instead of GNSS. This leads to an accumulation of many points with the exact same coordinate. These are then combined into a cluster which has the size zero. These small clusters were removed as well as all outliers. 85 clusters remained after these operations.

Another clustering operation was performed to find patterns within the larger clusters. As maintenance facilities are often located in large freight yards, this step was an attempt at separating the maintenance facilities from loading and coupling areas. A lower minimum number of points (10) and a smaller search distance were used (50 m) to adapt to the new cluster requirements. After again removing outliers and small clusters, 42 remained in Sweden and 78 in Italy. The convex hull was calculated for each cluster resulting in telemetry geofences (TGF). Figure 8 illustrates the TGF creation process in Malmö, Sweden. Appendix C shows the ArcGIS Pro models created to perform the above-described tasks.

### 3.3.3 Combination of infrastructure and GNSS geofences

Next, the information of both the infrastructure and the telemetry geofences were combined. Several assumptions were used in this process. Firstly, for one damage location with both a TGF and an IGF, telematics data was deemed more reliable than the address data. In larger train yards the maintenance buildings can be away from the street and the address geolocation might not be a good fit. The telematics data indicates that wagons were in a certain part of a trainyard while they were damaged. There are, however, locations other than a maintenance facility a wagon stays at while being out of use. Such railway sidings can be difficult to distinguish from maintenance rails and buildings. The second assumption entails that it is more appropriate for the functionality of the final algorithm to have rather larger geofences than smaller ones. For some damage locations, only very few damage protocols could be obtained for 2020. It is possible that the GNSS track varies for other damage cases. If the geofence is very small, relevant geofence entries might not be detected.

Different cases of IGF and TGF combinations can occur at different damage locations. The following scenarios are possible for one damage location:

- A) Both IGFs and TGFs present at the damage location and they overlap.
- B) Both IGFs and TGFs present at the damage location and they do not overlap.
- C) Only IGFs present at the damage location.
- D) Only TGFs present at the damage location.

If the IGFs and the TGFs overlap to a high degree (case A), it is very likely that the maintenance facility is within those geofences. Other IGFs in that area that do not overlap with TGFs, can be deemed as irrelevant as the damage location was already identified. Accordingly, IGFs and TGFs that overlap more than 40 % (relative area overlap of either polygon) were combined using the convex hull following the first assumption. If IGFs and TGFs overlapped more than

10 % and no other overlap was present at that damage location, the geofences were also combined. If other IGFs that do not overlap with an TGF were present within 5 km of the combined geofences, they were deleted. As it is uncommon to have several maintenance facilities in the same location, no other maintenance facility should be within 5 km of the found location. TGFs were kept as these areas were relevant for damaged wagons and different repair and movement pattern can occur during a damage, following the second assumption.

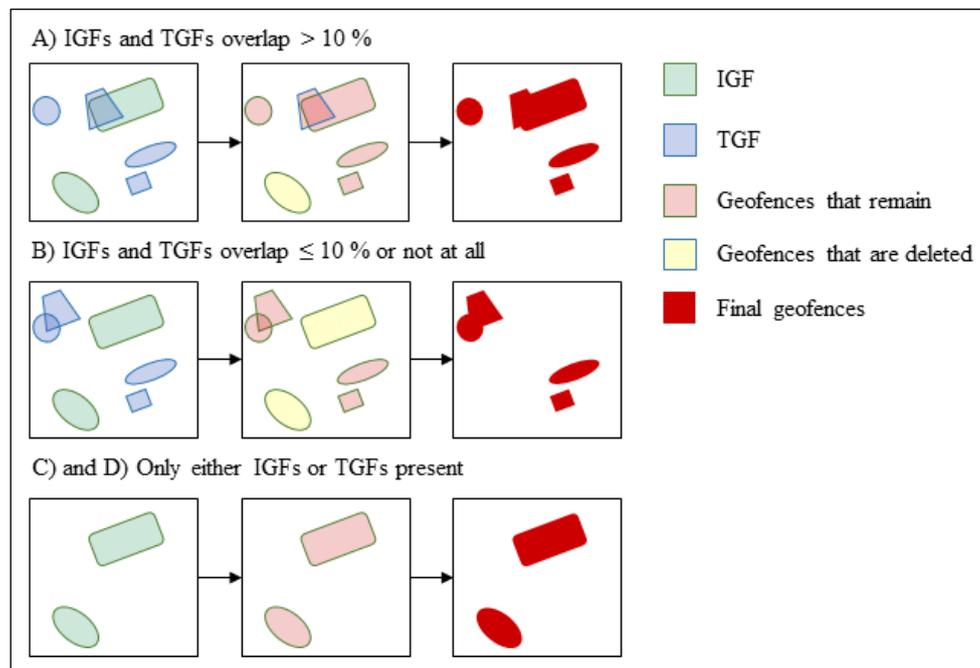


Figure 9: Cases for combination of IGFs and TGFs and process for each case.

If the IGFs and TGFs overlap less than 10 % or not at all (case B), the TGFs are kept, but the IGFs deleted. This is in accordance with assumption one as stated above. If only either IGFs or TGFs are present (cases C and D), they are kept. After deleting all irrelevant geofences, the remaining were combined using convex hull if they were within 50 m of another following assumption two. As a buffer of 10 m was already added while creating the geofences, no additional buffer was added. The different cases and the steps performed to combine the IGFs and the TGFs can be seen in Figure 9. Appendix D shows some of the more detailed ArcGIS Pro models used for this purpose.

### 3.4 Development of algorithm for geofence event detection

To comply with the aims set for this project, the developed algorithm must give the damage date, geofence entries and exits, and a substitute operations release (SOR) indicating when the repair of a wagon was finished at the latest. Additionally, it was desirable to include a quality indicator for each result which illustrates its reliability. To cope with the large data amounts, the same tools were used as for the telematics data preprocessing described in section 3.3.2. The algorithm is divided into six main sections: (1) setup and data import, (2) data preprocessing, (3) detecting geofence entries and exits, (4) filtering out relevant geofence

entries and exits, (5) computing operations releases, and (6) export. Figure 10 summarizes the steps the algorithm performs. For simplification purposes, the entry and exit of one wagon to one geofence will be referred to as one geofence event. The algorithm was iteratively created, adding functionality step by step. This way, intermediate results could be assessed and their quality improved. While testing, the data could be understood better, and special cases were detected and handled.

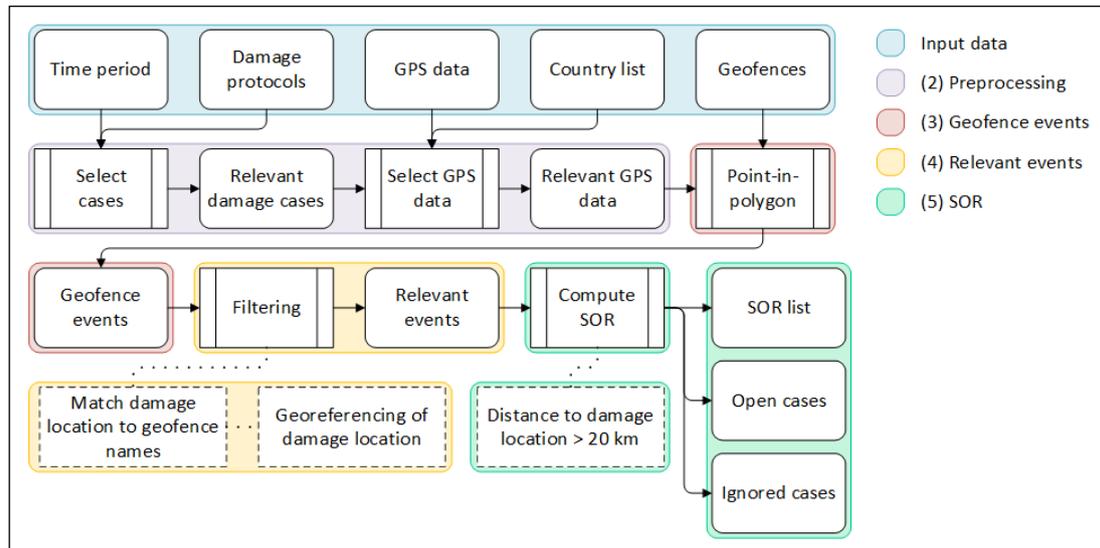


Figure 10: Most important steps performed in algorithm to derive geofence entries and exits and substitute operations release.

The inputs to the algorithm are comprised of the geofence data, WADIS damage protocol data, wagon telematics data, a time period, and a list of countries to consider. After the setup, WADIS damage data is loaded in the specified interval. For each damage case, the relevant GNSS data is loaded. If an OR is deposited in WADIS, the wagon data between the damage date and the WADIS OR is selected. If no OR is found, data until 30 days after the end of the user-given time period is loaded. This limitation was used to reduce the data amount to be processed. The geofence and telematics data are converted to geometry objects. To detect geofence events in step (3), each wagon GNSS point coordinate needed to be compared to the 113 geofences coordinates to determine if the point was within any of the geofences. To speed up this process, an anti-geofence was created including all areas in Europe outside of the maintenance geofences. Each point is first checked against this anti-geofence, and only if the point does not intersect it, the maintenance geofence it lies within is determined. To account for GNSS inaccuracies, a two-point logic was used to determine geofence entries and exits. This means that two consecutive GNSS points need to be within the same geofence to create an entry event. Likewise, an exit only gets generated if two consecutive points are outside of that geofence. The result of this step is a list of all geofence events for the damage cases in the selected time period with entry and exit time, and the length of the stay in it. As wagons can move to other geofences after being repaired, not all of these geofence events are relevant for the repair process, however.

To select the relevant events, connections between damage locations and geofences needed to be established. This follows the assumption that damaged wagons are repaired close to their damage location. The aim was to tag each maintenance geofence with damage locations which are likely to use this geofence for repair. Two approaches were used to realize this task. First, a string comparison was applied. As the WADIS damage location is not georeferenced, the use of geographic tools is limited. The closest railway station file, which states the closest station for each point in Europe using a Voronoi diagram, was used to transfer station names to close maintenance geofences. This file is also used for the closest railway station field in the wagon GNSS data. The centroid of each geofence was reverse-georeferenced to retrieve an additional relevant location name. This way, a geofence in Malmö, which so far only had a number, was assigned tags like Malmö Godsbangård, Malmö Frihamn, Malmö Central, and Malmö. However, not the closest rail station but the damage location is of interest. Therefore, a string comparison like the one in section 3.3.2 was performed between the rail station names associated to a geofence and the damage locations.

In a second approach, the WADIS damage locations were georeferenced using Google Maps. This was only possible as the country of the damage was added to the WADIS data export upon request. During the GNSS point selection in the geofence creation this method could not be applied as the data was not available yet. If a damage location was within 10 km of a geofence, it was deemed relevant for that geofence. The results of the two methods were assessed and compared. As the string comparison sometimes returned geographically completely wrong results, the georeferenced data was preferred while selecting the final geofence name tags. All damage cases existing in WADIS were used for these steps to get the maximum amount of relevant name tags. This part was performed only once in a separate preprocessing script.

The name tags were used to select relevant geofences by again performing a string comparison between the damage location and the name tags of a geofence the wagon entered. If the location strings matched, the geofence event was considered relevant. If no relevant event was found, the name tags of geofences close to the entered geofence were checked as well. If still no relevant event was found, the damage location was georeferenced and the proximity to geofences calculated. This way, new damage locations not included in the name tagging process described above, could be recognized as well. As georeferencing locations takes time and cost money if many queries are made, it was desirable to limit the amount of damage locations to be referenced as much as possible.

Lastly, the substitute operations release (SOR) indicating when a wagon left the maintenance facility area was computed. The relevant events needed to be filtered further to extract the last event connected to the repair. If a wagon moved further than 20 km away after leaving a relevant geofence, it was considered to be repaired and back in service. As wagons can be moved considerable distances on freight yards, this distance should imply a drive away from the maintenance facility and a completed repair. Additional geofence events after this exit were deemed unimportant. If a wagon is within a geofence, then for example moves 30 km away and

comes back to the same geofence afterwards, the second geofence entry will not be considered anymore. The date of the SOR is the day of the exit out of the last relevant geofence. If a wagon has not yet moved away far enough from the geofence, it is not considered to be repaired. These “open cases” are saved in a separate list and can be tracked in future periods. Cases which are neither in the open cases nor in the SOR list and do not have a WADIS OR are added to the ignored cases list. For these damage cases, no SOR could be generated.

### 3.5 Evaluation and validation

The quality of the results was tested and evaluated with different methods. The geofences were assessed by asking maintenance facility employees on the locations of maintenance facilities. Two employees from DB Italia and Duroc Rail AB gave the exact positions of buildings and rails belonging to their maintenance facilities. As some companies did not respond to inquiries, the information on cities with maintenance facilities available on their websites was used.

The geofence events and SOR were manually verified by plotting the wagon movement over time and comparing it to the events. Cases that did not have any events or SOR, were examined thoroughly to detect the cause for the algorithm failure. The quality of the substitute operations release (SOR) was additionally verified by comparing it to the OR deposited in WADIS. The SOR was assessed for damage cases with an OR in WADIS for damage dates between 01.04. and 30.06.2021. A total of 123 cases was analyzed. This time period was chosen as data from January to March 2021 was used to evaluate and improve the algorithm during the development phase. Almost no cases had ORs in WADIS if the damage date was later than June at the time of the analysis. As the final algorithm uses the WADIS OR if the SOR date is later, the algorithm was adapted so that damage cases with WADIS ORs were treated as if they had no OR. Basic statistical values like standard deviation and mean were computed to assess the difference in SOR and WADIS OR.

Further, a quality indicator  $Q$  was included in the algorithm, giving an indication through which means the SOR was created. It combines an indicator  $Q_{GF}$  for the geofence quality and an indicator  $Q_{SOR}$  for the SOR quality. For  $Q_{GF}$  four different categories were defined. For  $Q_{SOR}$ , three different categories were defined. They are described in Table 3. The values for the geofence quality and the SOR quality were summed up.  $Q$  has a range between 0 and 4. A low  $Q$  indicates a higher reliability of the result than a higher  $Q$ .

Table 3: Definition of quality indicators  $Q_{GF}$  and  $Q_{SOR}$ .

Value	Cases in which value is assigned
<b>Quality indicator for geofences <math>Q_{GF}</math></b>	
0	IGF and TGF overlap more than 10 % or Geofence is a TGF with overlapping IGF and TGF close by
1	Geofence is a TGF, other IGF close by
2	Geofence is a TGF, no other IGF close by
3	Geofence is an IGF, no other TGF close by
<b>Quality indicator for SOR result <math>Q_{SOR}</math></b>	
0	Wagon is more than 20 km away from last relevant geofence
1	Case with an OR in WADIS, SOR as good or worse than WADIS OR
2	Case with an OR in WADIS, no SOR generated, WADIS OR used

To answer the 3<sup>rd</sup> research question, three geofence categories were introduced (Figure 11). They depict the most common relative geofence locations. Category A is a geofence around a maintenance facility that is separate from the main rail network. Usually one single rail would lead to the area where the maintenance facility is located. Category B geofences are located within large freight yards where more than five rails are located next to each other. Geofences in category C are located close to the main tracks with up to five parallel rails.

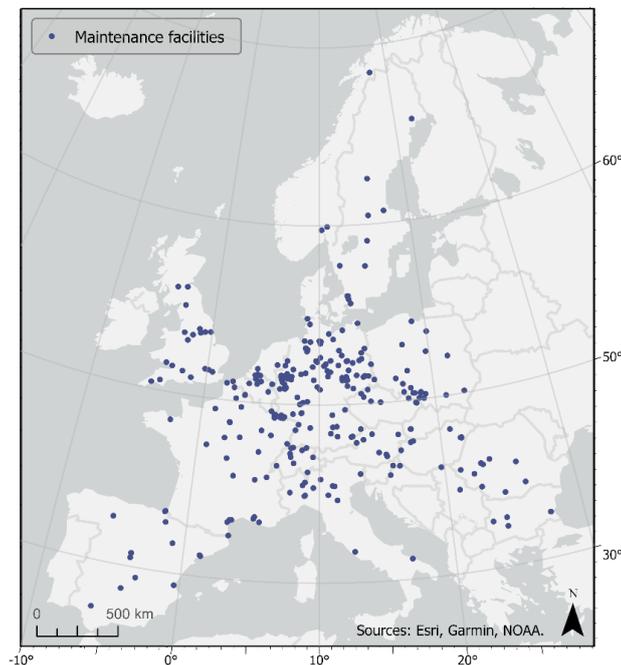


Figure 11: Examples of geofence categories. A) Category A: Geofence separate from main rail network. B) Category B: Geofence at freight yard. C) Category C: Geofence close to main rails. Sources: Esri, HERE, Garmin, Lantmäteriet, USGS, OpenStreetMap.

## 4. Results

### 4.1 Geofence creation

During the geofence creation, 120 IGFs and 120 TGFs were computed for Italy and Sweden. That the number of IGFs and TGFs is the same, is a coincidence. After the combination of the IGFs and TGFs, 113 geofences remained, 37 in Sweden and 76 in Italy. These geofences are located at 17 locations in Sweden and 27 in Italy. The maintenance facility locations can be seen in Figure 12 and Figure 13. Two of the geofence locations in Italy are not directly in Italy but in a neighboring country at the Italian border. One is located in Chiasso, Switzerland and the other in Dobova, Slovenia. Both are TGFs and their creation can be attributed to the closest rail station associated to the telematics data. During the creation of the closest rail station Voronoi diagram at DB Cargo, country borders were not considered. Thus, one closest rail station cell can cross borders. Only one of the two countries that the cell intersects is added to the metadata. In Chiasso and Dobova, the metadata connected to the wagon states that the wagon is still in Italy. This is a data selection issue and not an inherent problem of the proposed geofence creation methodology. Thus, as the created geofences are not necessarily incorrect, they were kept in the dataset.



*Figure 12: Maintenance facilities in Europe identified through DB and VPI address data. Total number: 311. Projection: EPSG:3035.*

The geofence locations were cross-referenced to the maintenance facility addresses. Out of the 14 address points in Italy, 5 were within 10 km of a TGF. The same was true for 6 out of the 12 Swedish maintenance facility addresses. In Sweden, for all address points an IGF was created nearby. In Italy, two address points had no IGF associated to them. Those two locations

were also not close to any TGFs. One of those addresses is in Caldiero close to Verona, where another address point with geofences is located. The other is in Lovere far off from any rail infrastructure. Two examples for the created geofences can be seen in Figure 14 for Malmö and Rivalta Scrivia.

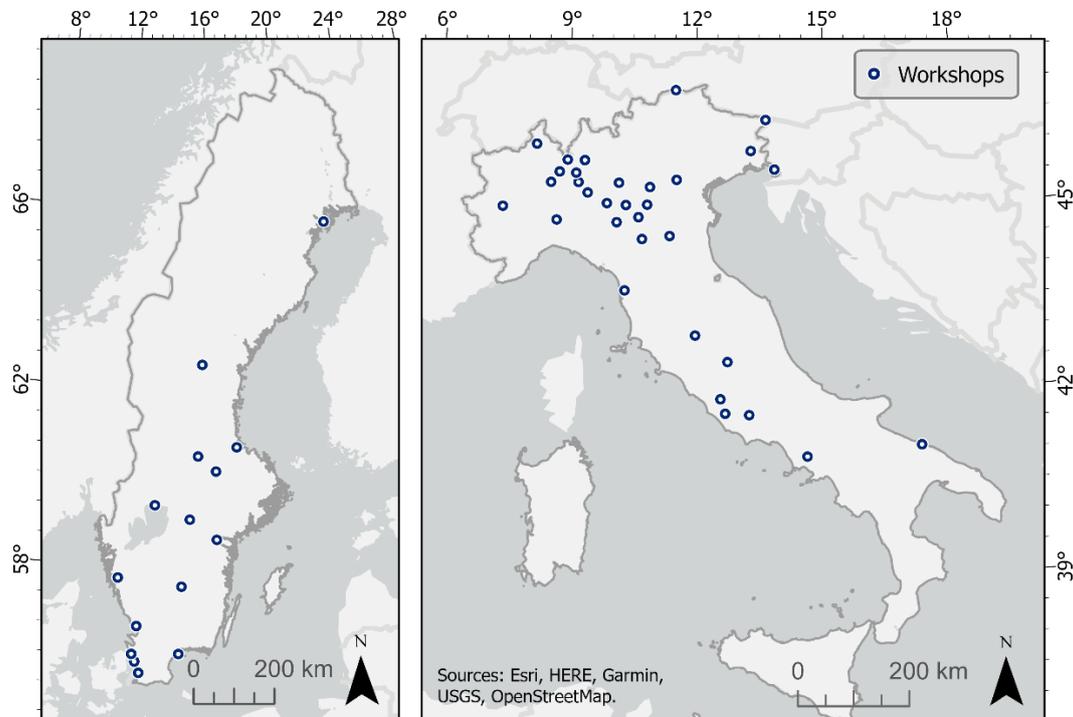


Figure 13: Locations in Sweden and Italy where the created geofences are situated. Close maintenance facilities are aggregated and represented as one point. Projection: ETRS89 LAEA.

A total of 23 of the 44 geofence locations could be completely or partially verified through maintenance employee feedback and maintenance provider information. 9 of the 23 geofence locations were validated using direct feedback from maintenance employees, one in Sweden and eight in Italy. The other 14 geofence locations were validated partially by comparing the stated maintenance facility locations on Swedish maintenance provider websites.

One maintenance provider in Sweden (Duroc Rail AB) and one RC in Italy (DB Cargo Italia) responded to inquiries on their respective maintenance facility locations. Out of the nine geofence locations checked through direct feedback of maintenance employees, four fully matched the actual maintenance facility location. Two overlapped the actual location partly, two were located close to the actual building and one was at a location where no maintenance facility exists. Often, the geofence is larger than actual maintenance facility. There is no indication that IGFs or TGFs are more reliable than the other. Of the nine maintenance locations considered, five only had TGFs, three only had IGFs and one had both IGFs and TGFs.

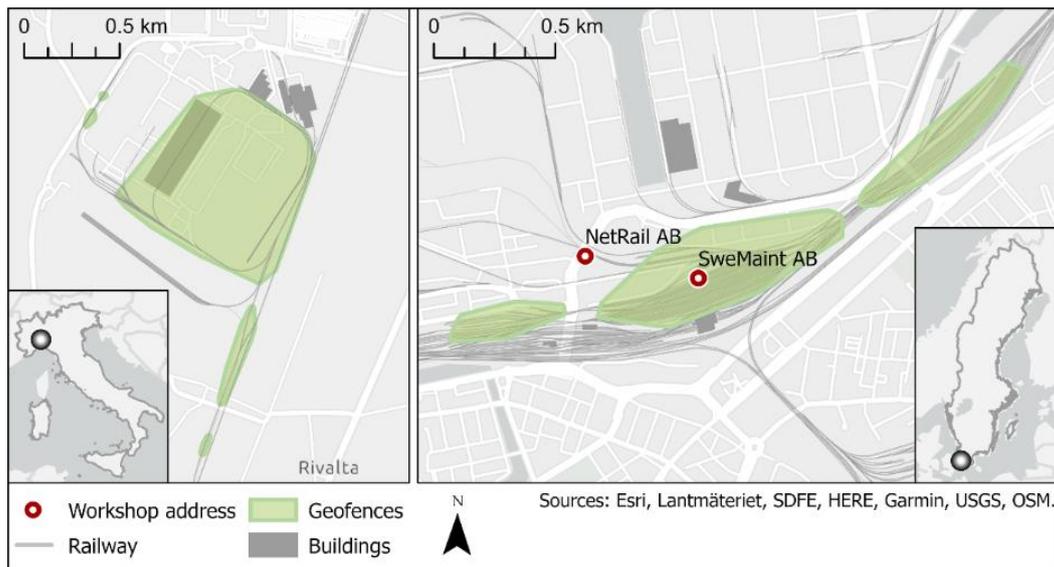


Figure 14: Examples of final geofences in Italy and Sweden. Left: Rivalta Scrivia. Right: Malmö.

Figure 15 shows an example for feedback from Duroc Rail AB on the maintenance facility geofences in Luleå. There, the geofence created through this project is only comprised of IGFs. As shown in the figure, the address georeferencing worked reasonably well although the actual location is situated approximately 780 m from the address point. The created (green) geofence partly covers the actual maintenance facility (red). The maintenance facility is comprised of one main access rail and one building. This building is not included in the OSM database, however. If that would have been the case, a greater amount of the actual maintenance facility might have been covered.

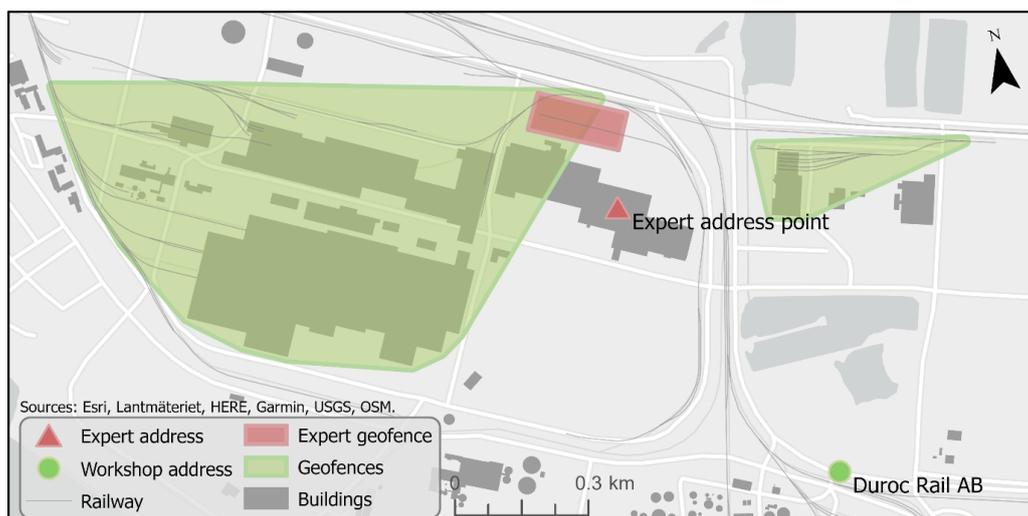


Figure 15: Example for feedback from Duroc Rail AB on maintenance location in Luleå. The created maintenance geofence (green) can be compared to the actual location of the maintenance facility (red).

Additionally to the feedback from rail maintenance workers, 14 cities with maintenance facilities in Sweden could be confirmed through information stated on SweMaint's and NetRail's websites, two large maintenance providers in Sweden (NetRail AB 2021; SweMaint

AB 2021). At 7 of the 12 maintenance facility locations of SweMaint maintenance geofences are present. 7 of the 12 also had address points. The address and connected geofence of Gävle seem to be faulty however, as the wagons stay at a different location outside of the geofence. Furthermore, one address point in Helsingborg was incorrect, as there is no maintenance facility there. Both maintenance locations of NetRail have addresses and geofences associated to them.

To summarize, the proposed methodology of using address and infrastructure data on the one hand, and GNSS data on the other hand and combining the result worked. The created geofences depict actual maintenance facility locations reasonably well. Most existing maintenance facilities are included in the result. Often, the geofence is larger than the actual maintenance area. The combination of IGFs and TGFs provide the most reliable results. Both faulty addresses and excess TGFs decrease the quality of the result.

4.2 Algorithm development

Matching relevant damage location to geofences was an important component of creating a reliable algorithm. The string comparison returned matches for 109 of the 113 geofences. The quality was not always sufficient, however. Words that are rail-specific and not connected to a location, e.g. godsbangård in Sweden, impair the quality of the result. In other cases, names of cities are too similar. The consequence is, that damage locations were matched to very distant geofences. Georeferencing the WADIS damage location with a Google Maps API returned more reliable results. Damage location could be matched to 106 geofences this way. The final location tags include 247 out of 446 unique damage locations in Italy and Sweden in 2020. Although this means that only 55 % of damage location in this period were matched, these cover over 90 % of all damage cases in 2020. The most important damage locations were assigned, therefore. Each geofence has an average of 5.7 damage locations assigned to it. Often, they include the same name with different spellings. Missing locations were handled by integrating damage location georeferencing and geofence distance calculation in the algorithm.

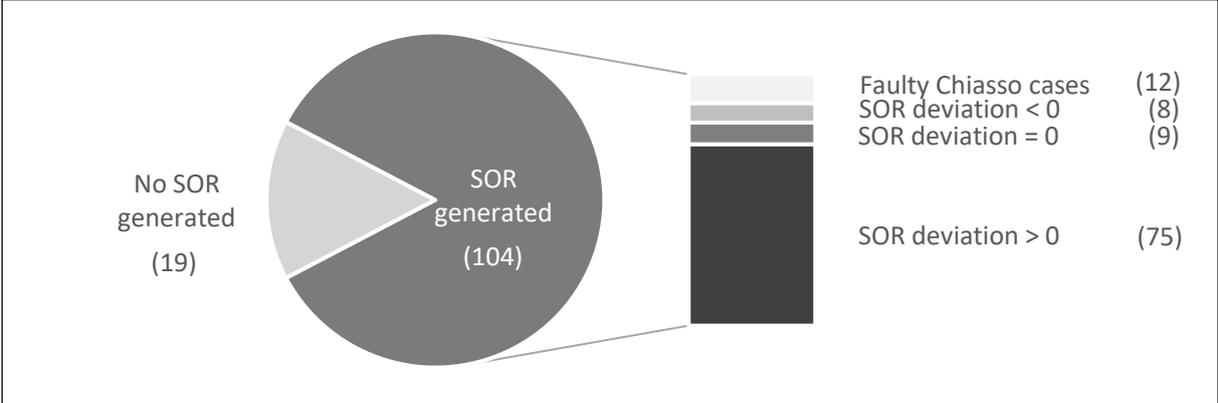


Figure 16: Overview of effectiveness of algorithm in calculating substitute operations releases (SOR). Total sample size: 123. The brackets show the number of cases in this category. The Chiasso cases had faulty GNSS data and were not included in the statistical analysis. The SOR deviation results from deducting the actual OR day from the SOR the algorithm generated. A positive SOR deviation indicates that a wagon stayed at the repair location beyond the OR date while a negative SOR deviation means the wagon left the repair location before the OR date.

To evaluate the efficiency of the algorithm, the SOR was calculated for damage cases with an OR in WADIS for damage dates between 01.04. and 30.06.2021. A total of 123 cases was analyzed. For 104 of the 123 cases (86.6 %), a SOR was generated. The reasons for the missing SOR were no geofences at the damage locations, too small geofences, or faulty GNSS data. 8 of the 19 cases without a SOR had damage locations in Italy, the remaining 11 in Sweden. For the 104 cases with a SOR, the difference to the date of the actual OR in WADIS was calculated. As the OR date in WADIS can be faulty, as discussed before, the OR in the actual OR document was used. If there was a mismatch, the reason for the difference was examined using the GNSS tracks and damage data. Figure 16 gives an overview of the algorithm results for the sample data. For 12 of the 104 cases, the SOR was considerably earlier than the actual OR. An investigation of those cases showed that they all happened in Chiasso, Switzerland at the Italian border. Although the wagon stayed in Chiasso throughout the entirety of the damage, an earlier SOR date was computed. As only data for Italy and Sweden was used in the algorithm (wagon data with closest railway station that has the country code of Italy or Sweden), GNSS data a bit further away from the border was excluded. Therefore, not all relevant GNSS points for these damages were included. These cases were ignored in the statistical evaluation.

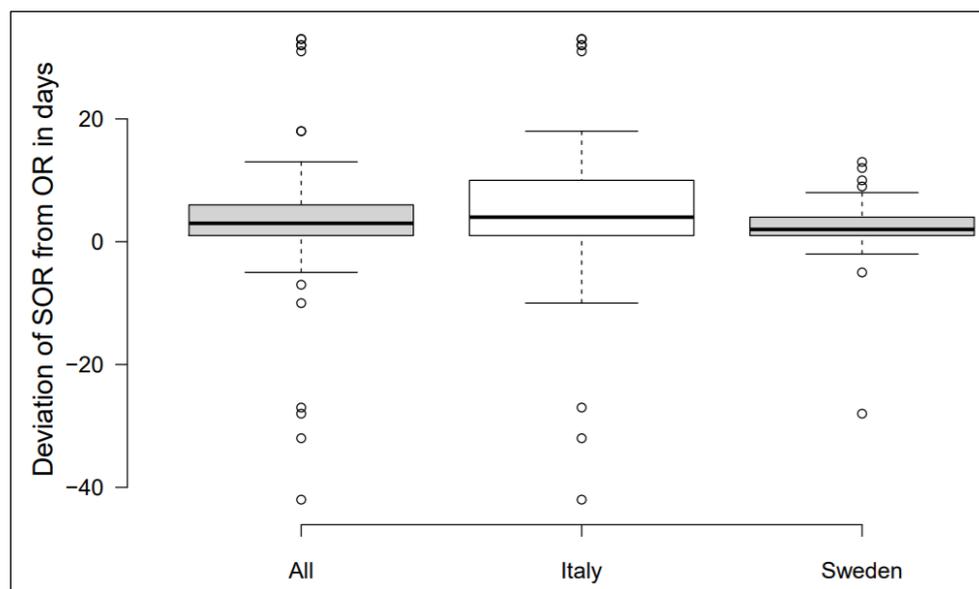


Figure 17: Distribution of deviation in SOR date from actual OR date in days for damage cases from 01.04.21 to 30.06.2021 in Italy and Sweden. Sample size: 92 damage cases. The associated statistical values can be seen in Table 4.

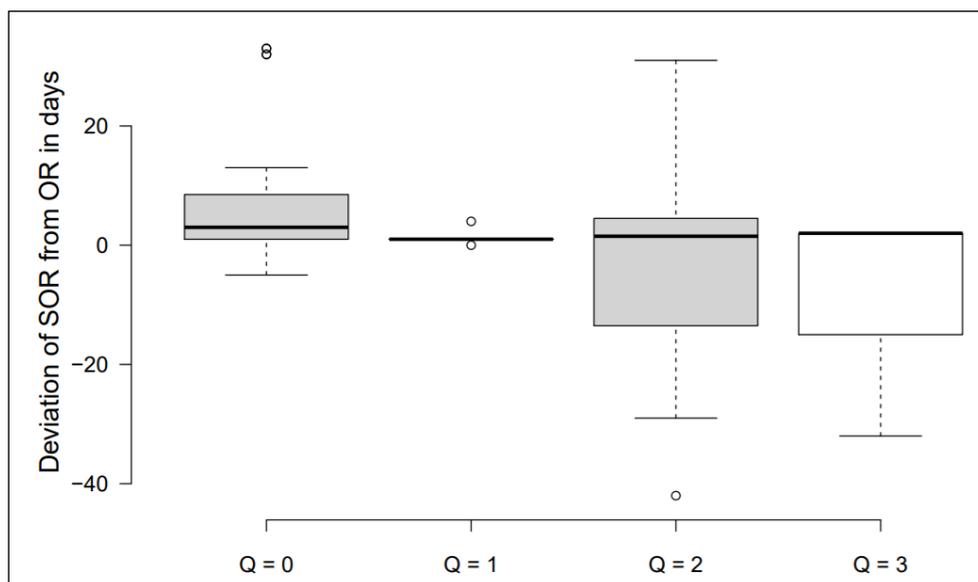
Figure 17 shows the distribution of the SOR deviation for all 92 remaining cases, and for damage cases in Italy and Sweden separately. Table 4 gives the associated statistical values. The median difference between the SOR and the actual OR amounts to plus three days for the entire sample. The mean is almost four days. This means that the detected latest relevant geofence exit (SOR) was on average four days later than the actual OR. With a standard error of the mean (SEM) of 1.2 days, a 95 % confidence interval of [2.5 days, 4.9 days] can be derived. In most cases (75 of 92), the SOR was after the actual OR. This is caused by a stay of

a wagon at a damage location beyond the repair. In 9 of the 92 cases (8.7 %) the SOR had exactly the same date as the OR in WADIS.

*Table 4: Statistical values for boxplots in Figure 17, giving the SOR deviation for the entire dataset and disaggregated for each Italy and Sweden. The unit for all values except the number of cases is days.*

	All	Italy	Sweden
Upper whisker	13.0	18.0	8.0
3rd quartile	6.0	10.0	4.0
Median	3.0	4.0	2.0
1st quartile	1.0	1.0	1.0
Lower whisker	-5.0	-10.0	-2.0
Mean	3.7	5.4	2.3
Standard deviation	11.1	15.1	5.5
Standard error of the mean	1.2	2.3	0.8
Number of cases	92	42	50

The SOR for damage cases in Sweden is more accurate than in Italy. The mean SOR deviation in Sweden is only about two days, while it is more than five days in Italy. The SEM and the standard deviation are also lower in Sweden than in Italy. Some Italian cases were far off the actual OR date while Sweden's outliers are less pronounced. Several damage cases in Rivalta Scrivia, Italy had a much later SOR date because the wagon stayed at the damage location long after the repair was finished. In other cases in Chiasso, the wagon moved over 100 km away from the damage location while being damaged according to WADIS. That lead to a much earlier repair date.



*Figure 18: Distribution of deviation in SOR date from actual OR date in days for damage cases disaggregated according to quality indicator. A low  $Q$  value should indicate a higher SOR accuracy while a higher  $Q$  value should indicate lower SOR accuracy. Time period: 01.04.21 to 30.06.2021. Sample size: 92 damage cases. The associated statistical values can be seen in Table 5.*

The SOR deviation disaggregated according to the combined quality indicator  $Q$  that describes the quality of the algorithm output can be seen in Figure 18 and the corresponding statistical values in Table 5 (for explanation of  $Q$  see section 3.5). The data for  $Q$  equaling one and three can be ignored in the evaluation as only very few cases had these  $Q$  values. There seems to be a weak link between  $Q$  and the SOR deviation. Although the median and mean value are closer to zero for  $Q=2$  than for  $Q=0$ , the variance is much higher. This means that even though that cases with  $Q=2$  have a SOR value closer to the actual OR date on average, the probability of having a SOR that is very different from the OR is much higher with this quality indicator. The assumption that a higher  $Q$  would give worse results was at least partly confirmed, therefore.

*Table 5: Statistical values for boxplots in Figure 18, SOR deviation disaggregated according to quality indicator  $Q$ . The unit for all values except the number of cases is days.*

	$Q = 0$	$Q = 1$	$Q = 2$	$Q = 3$
Upper whisker	13.0	1.0	31.0	2.0
3rd quartile	8.5	1.0	4.5	2.0
Median	3.0	1.0	1.5	2.0
1st quartile	1.0	1.0	-13.5	-15.0
Lower whisker	-5.0	1.0	-29.0	-32.0
Mean	6.4	1.3	-3.2	-9.3
Standard deviation	9.1	1.4	14.1	19.6
Standard error of the mean	1.3	0.6	2.0	11.3
Number of cases	47	6	48	3

To answer the third research question, the SOR deviation needed to be disaggregated according to the geofence categories. Each geofence was assigned to a category resulting in 40 geofences in category A, 53 in category B and 20 geofences in category C (for the category definition see section 3.5). The category of the last relevant geofence a wagon exited was used. None of the 92 cases last exited a category A geofence that is separated from the main rail network. Thus, only categories B and C can be evaluated. The distribution of the SOR deviation is shown in Figure 19, the statistical values in Table 6.

B and C are similar categories. For both, the maintenance geofence is not on a separate rail as in category A, but it is close to freight yards (category B) or to main rail routes (category C). As most geofences are category B geofences, more sample cases were last in a category B geofence. It can be assumed that the low sample size of category C cases increased the standard deviation of the dataset. With almost 17 days it is much higher than the standard deviation of category B with 9 days. The mean and median SOR deviation is also considerably higher for category C. The mean and median SOR deviation of category B are both approximately 2 days. They are closer to zero than the equivalent values for the entire dataset. It is surprising that the category B geofences produced such good results. It was expected that geofences at freight yards would yield worse results as it is more likely that a wagon stays there longer after the repair to await a new disposition order. It is very likely that the uneven sample size has a considerate effect on the result, however.

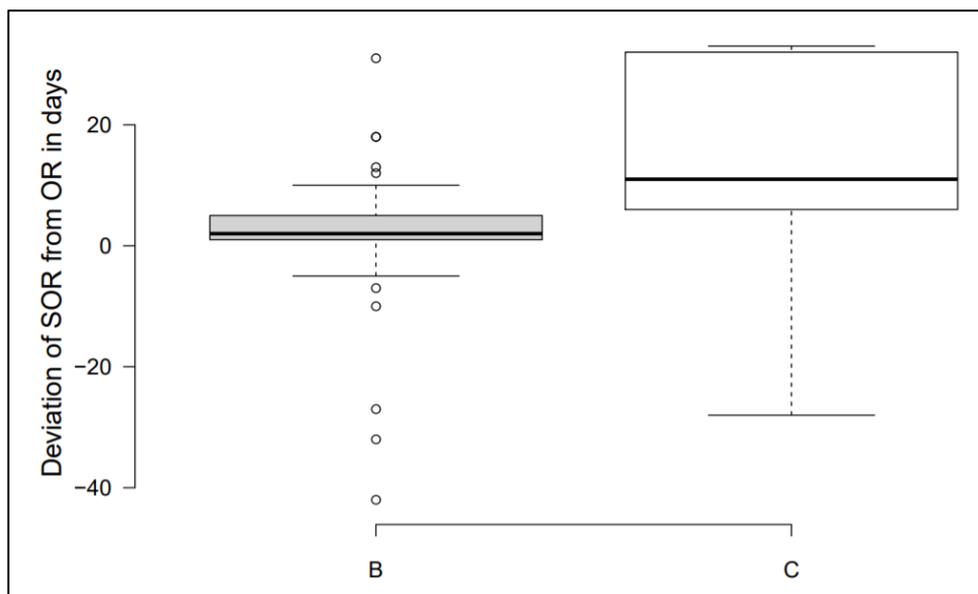


Figure 19: Distribution of deviation in SOR date from actual OR date in days for damage cases disaggregated according to the category of the last relevant geofence exited. Category B: The maintenance geofence is close to or within a freight yard. Category C: The maintenance geofence is close to main railway routes. Time period: 01.04.21 to 30.06.2021. Sample size: 92 damage cases. The associated statistical values can be seen in Table 6.

To summarize, the algorithm produced an SOR for over 86 % of the sample damage cases. On average, the SOR was approximately 4 days after the OR in WADIS. In most cases, the wagon thus stayed at the damage location after the repair had been finished. No clear findings can be derived from the quality indicator  $Q$  as two of the four values could not be evaluated. The same is partly true for the SOR quality of the different geofence categories as no geofences separate from the main rails (category A) could be analyzed. The comparatively high accuracy of geofences at freight yards (category B) was unexpected.

Table 6: Statistical values for boxplots in Figure 19, SOR deviation disaggregated according to geofence category. Category B: The maintenance geofence is close to or within a freight yard. Category C: The maintenance geofence is close to main railway routes. The unit for all values except the number of cases is days.

Category	B	C
Upper whisker	10.0	33.0
3rd quartile	5.0	32.0
Median	2.0	11.0
1st quartile	1.0	6.0
Lower whisker	-5.0	-28.0
Mean	2.1	13.2
Standard deviation	9.0	16.9
Standard error of the mean	1.0	4.7
Number of cases	79	13

## 5. Discussion

### 5.1 Interpretation of results

In the following section, the above stated results will be interpreted and evaluated. It is not possible to conclusively rate the quality of the created maintenance geofences. 314 addresses of rail maintenance facilities were collected. To assess its completeness, a list of company names that performed repairs on DB Cargo wagons in the past was obtained. That list has 310 entries which is very close to the total of 314 obtained with the VPI and DB data. However, as the list only contains company names and not their respective locations, it is likely that the actual number of maintenance facilities is higher as one maintenance company can have several locations. But as the range is similar, the address list should cover most maintenance facilities.

The TGFs were used to find the missing maintenance facilities. This worked well in some cases like Verona, Chiasso, and Avesta. As no TGFs were deleted, it is likely that there are TGF locations without an actual maintenance facility. The assumption that every wagon is repaired at its damage location is not valid for every damage case, although for most. Cases have been found where the repair location was unknown. One wagon with a damage location in Malmö only stopped there for a couple of hours and then continued to Borlänge some hundred kilometers away. According to WADIS, the OR was issued in Borlänge. Did the maintenance facility in Malmö assess the wagon but it was more sensible to repair it in Borlänge? Are there damages that cannot be repaired in Malmö? Was the OR date wrong? There are several possibilities and the actual reason cannot be determined with the available data.

The number of combined geofences in Italy and Sweden is slightly higher than the number of address points in those countries (Sweden: 12 address points, 17 geofence locations; Italy: 16 address points, 27 geofence locations). These results are in accordance with the assumption that the address list was incomplete and additional maintenance locations needed to be identified with telematics data. Without further feedback from maintenance providers, it cannot be known which of the new locations really have a maintenance facility, however.

Another task of the TGFs was to validate the IGFs. This worked well for example in Rivalta Scrivia, Guastalla (both Italy) and Ånge (Sweden). In many cases the TGF was larger than the IGF, e.g. Malmö and Nässjö. This is also due to the choice of making geofences rather larger instead of risking too small geofences without any entries. The results of the algorithm support this decision. Too small geofences were the reason for no SOR in some cases like Karlstad. Although the result is less precise with larger geofences and it is more unlikely that freight yards will be separated from maintenance facilities, it is more desirable to get any SOR than none. The larger geofences can also be one cause of the delay in the SOR, especially in category B and C geofences where wagons tend to stay longer after a repair to await new disposition orders or to get assembled to a train.

Although the algorithm includes a georeferencing function that is used if no relevant geofence can be found, the assignment of relevant WADIS damage location to each geofence is still important. The georeferencing of the damage locations worked well once the country was added to the WADIS export. Without a country reference, the result was often incorrect. The availability of this data during the geofence creation phase could have improved the selection of relevant GNSS data points and the clustering.

The algorithm successfully created SORs in more than 86 % of damage cases in the selected period. This is a very satisfactory result as it creates knowledge about the repair status of most wagons while abroad. Given that currently, no OR is sent at all in most cases, an average deviation of three or four days from the actual OR is acceptable. The difference in SOR quality between Sweden and Italy was unexpected. During the testing phase of the algorithm, more problems occurred with Swedish damage locations. Although Sweden has more cases where no SOR could be produced, especially when taking the total number of damage cases per country into account, the quality of the created ORs is distinctively higher than in Italy.

While the quality indicator  $Q$  shows a higher variance at a value of two than of zero, the median and mean of  $Q=2$  is closer to zero, indicating a better SOR on average. The explanatory power of  $Q$  remains limited, therefore. With more data with  $Q$  equaling one and three, a clearer trend might be observed. Insufficient data also limits the evaluation of the difference in SOR for the three geofence categories. It was expected that category A would be decidedly better than categories B and C. As the geofence location in the two latter categories is relatively similar, no large difference in SOR quality was expected. The observed superiority of category B over category C geofences was unforeseen. As the standard deviation is high in both cases, and the SEM is still very high for C, the robustness of the result is limited, however. It is very unfortunate that no category A geofences were last exited in the analyzed period. A clearer trend might have been detected otherwise.

## 5.2 Reflection on validity, objectivity, and reliability of the results

Developing a new methodology is always connected to uncertainty about its effectiveness. Often, the researcher has to choose one options out of several acceptable ones without knowing for certain how it will affect the output. The preferred value or process needs to be justified rigorously. This step can be demanding, however, especially when it comes to parameters that cannot be chosen based on scientific knowledge. Why was a distance of 20 m and not 19 m or 15 m chosen? Including experience from local maintenance facility employees, systematically testing different parameter settings, and assessing the result are approaches that were used to handle these cases in this thesis. As time and resources are limited during a project, it is also task of the scientific community to further develop and test existing proposed methods to refine and improve them.

Wagon data from damage cases in 2020 was used for the clustering in the TGF creation. The COVID-19 pandemic that started in early 2020 influenced all parts of society and economy including the freight rail sector. It is possible that the pandemic affected the proceedings of single damage cases, e.g. the time needed for a repair, to send spare parts, or to assign a new disposition order to a wagon. A time period without these confounding factors would have been preferred. However, the limited number of wagons equipped with telematics units before 2020 and the need to reserve data for evaluation left no other option. As the factors described above should mainly lead to longer repair times and longer stays at damage locations, the functionality of the algorithm or the shape of the created geofences should not be affected substantially.

Retrospectively, the 10 % threshold for combining the TGFs and IGFs seems too arbitrary. Some of those cases overlapped so little that it did not seem sensible to combine them. It might still have been a better choice to combine all overlapping geofences, however. Testing the algorithm with this change in the geofence creation methodology would show the impact of this decision. The addresses and derived IGFs played an important role in geofence quality. The approach of selecting relevant rail infrastructure worked well in many cases. Wrong addresses can considerably impair the result, however, as the geofence in Gävle showed. The IGFs could be further improved by including rails that are a bit further away from a maintenance building but connected and parallel to rails in the geofence selection. Also, implementing the functionality to select relevant rails if no building is close to an address point could be useful to detect open-air maintenance locations. As many TGFs were created, it is very likely that some of them are not at maintenance locations. The combination of IGFs and TGFs remains the most reliable result.

Two geofences were created outside of Italy and Sweden. Further, there were problems with damage cases in Chiasso, Switzerland during the SOR computation. These issues could have likely been avoided by using the damage country stated in WADIS to select GNSS data instead of the country stated in the telematics metadata. As this information was not available from the start, the clustering points were selected according to the country code associated to the closest train station in the wagon data. Yet, the country selection in the final algorithm should be performed through the now accessible WADIS damage country to avoid GNSS data issues.

The quality of WADIS damage day data is high. As most damages are due to abrasion and the wagon owner has to pay for abrasion repairs, it is in the interest of the RC using the wagon to report the damage and get reimbursed for the repair costs. Other parts of the WADIS data remain less accurate. Encoding problems while reading some damage protocols lead to faulty characters in the damage location, e.g. *Malmã Godsbangã...rd* instead of *Malmö Godbangård*. This can lead to problems while matching damage locations to geofences and selecting relevant geofence entries for a wagon. This problem only appeared until April 2021. In recent damage cases this complication did not occur. As the source of the issue was not identified, however, it could arise again in the future.

Wagons often remain at the damage location after the repair. Consultations with DB Cargo and DB Cargo Italia employees indicate that likely reasons for this are delays in communicating the updated status of a wagon, delays of the operating RC to pick up the wagon, or delays in assigning a new disposition orders to the wagon. As larger geofences were chosen, this behavior cannot be traced with the algorithm. When an SOR has been computed, however, it is highly likely that the wagon really is repaired and back in operation. Analyzing the difference in SOR and OR can also point out locations where wagons remain longer more often. A more detailed investigation of the processes at these locations can show efficiency deficits in handling repaired wagons which can then be addressed. This only applies to cases with ORs, however.

As only damage cases with ORs were analyzed in the sample period from April to June 2021, only a limited amount of damage locations was included in the analysis. It seems that some maintenance facilities issue ORs almost always while others never do so. 22 different damage locations were included in the sample. The algorithm has deficits when a damage location is very far from the next maintenance facility geofence. The assumption that more severe damages get repaired at the damage location would imply that there is a missing maintenance facility. There are some cases, however, where this assumption does not apply. Single cases where a wagon travelled to a distant maintenance facility were observed. Also, mobile maintenance is very common in remote regions. The algorithm cannot handle these cases. Most likely, it will not generate an SOR in these events.

### 5.3 Results within research context

Data availability and quality differs between different scientific disciplines and research contexts. In the cargo rail industry, there are still many applications where data is either not generated, properly stored, exchanged, or maintained. OSM also has deficits when it comes to detailed information about specific geographic objects. Maintenance facilities could not be identified by an OSM search. Other studies likely face similar problems. Rabiei-Dastjerdi et al. (2020) showed that even in London, where the OSM community is very active, facilities like pharmacies and swimming pools were not correctly tagged in the majority of cases.

Methods for extracting the needed data with the help of other available data are needed, therefore. This study proposes one approach that uses existing address and GNSS data. In combination with the existing OSM data, spatial analysis tools can be used to identify the areas of interest. Clustering played an important role in detecting repair locations. DBSCAN was the most suitable clustering method and handled the different cluster shapes and sizes well. This confirms Gülagiz and Sahin's (2017) results. The quality of the clustering process highly depends on the dataset, however. The data density needed to be limited in this study to enable a proper functioning of DBSCAN. Otherwise, nuances in denser clusters could not have been detected.

Geofences proved to be an adequate instrument to register wagon movements in and out of damage locations. For most damage cases, an SOR could be produced with an acceptable quality. As the creation of the geofences was accompanied by a limited data quality (e.g. in completeness of addresses and the selection of relevant GNSS points), larger geofences were desirable to compensate for uncertainties in the result. If the data allows to create highly accurate geofences only around the relevant infrastructure, smaller geofences can produce better results as freight yards can be better separated from maintenance buildings and rails. As the rails used for maintenance activities can change and GNSS errors occur regularly, it remains difficult to perfectly capture all relevant wagon movements even with smaller geofences, however.

Letting the data guide the research can work, as this study showed. Combining it with a hypothesis-driven approach helps guide the research activities in the desired direction. As time and money are limited resources, scientific research should use them meaningfully. Randomly extracting information from big datasets is not always producing useful knowledge. Human hypothesis, societal needs, and real-world applications can indicate which direction big data research should focus on. The usage of data-driven methods made it important to iteratively perform tasks and respond to intermediate results with additional steps if necessary. The quality of the results needed to be assessed continually and if needed improved.

GNSS data has only been available for a short amount of time in the cargo rail industry. This project demonstrated one application of the data. Many other uses are possible, however, e.g. in fleet management, route optimization, and predictive maintenance (e.g. Railnova 2021). As political prioritization and affordable technology foster innovation in the sector, further research and modernization are highly likely in the future. At the same time, customers require tracking functionality, data interoperability and a more efficient services to choose rail over road for transporting goods. Market forecasts from Grand View Research (2021) and Research and Markets (2021) predict that the rail industry will increase to rely on smart digital transportation systems like telematics units in the future. Rail companies in countries like Germany, India, Canada, and France are heavily investing in telematics units and other digital technologies (Grand View Research 2021). While the U.S. still has the largest rail telematics market, the European one is growing the fastest (Research and Markets 2021).

## 6. Conclusion

The aim of this thesis was to contribute new methodological approaches to the digitalization of the railway industry and to geographical information systems sciences. A methodology for the automated creation of geofences around existing but insufficiently referenced geographic objects was proposed. Based on address, infrastructure and GNSS wagon data, geofences around rail maintenance facilities in Sweden and Italy could be generated with an acceptable quality. Spatial analysis and clustering tools proved to be adequate instruments for this purpose. This gives another example how GIS can help to solve real-world problems. Creating infrastructure and telematics geofences separately and using them both to compute the final geofences allowed to increase the credibility of the result. It also laid the shortcomings of each approach open. IGFs rely on accurate addresses and complete buildings and rail infrastructure data. Errors in the input data quality will influence the result quality. TGFs are prone to give too many additional and possibly faulty maintenance locations as no process for effectively filtering TGFs was found. The automatically created geofences cannot fully replace the specific knowledge kept by maintenance providers about the exact location of a maintenance facility.

Next to the geofence methodology, an algorithm which detects geofence entries and exits and approximates the day a wagon was back in operation was developed. The goal of this part of the project was to give railway companies, in this case DB Cargo, a better insight into the repair status of their wagons while abroad. The up to now mostly missing operations release date could be generated for over 86 % of the analyzed damage cases. The deviation to the actual OR date amounts to +4 days on average. This is caused by a prolonged stay of a wagon at its repair location, and therefore the maintenance geofence, before it continues its current disposition order or is assigned a new order.

No clear relation between SOR quality and relative geofence location could be established. The analysis was impeded as no damage case in which a wagon last left a category A geofence (separated from the main tracks) was included in the examined dataset. Geofences in freight yards (category B) unexpectedly produced better SOR results than the entire dataset while wagons that last stayed in category C geofences (close to main tracks) delivered results with a high variance and deviation from the actual OR date. An uneven distribution of damage cases in each category limit the explanatory power of the geofence category, however.

The results of this thesis contribute to digitalization efforts in the cargo rail industry and test the application of GIS and spatial analysis tools in a new context. The proposed methodologies need to be tested and developed further, to increase the robustness of their results. Likewise, the creation of maintenance geofences should be extended to all of Europe to enable the computation of operations release dates for the entire European rail freight market.

## References

- Anderson, C. 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired* 16.
- Ballis, A., and L. Dimitriou. 2010. Issues on railway wagon asset management using advanced information systems. *Transportation Research Part C: Emerging Technologies* 18: 807–820. doi: 10.1016/j.trc.2009.09.003
- Balog, M., and M. Mindas. 2017. Informatization of Rail Freight Wagon by Implementation of the RFID Technology. In *Smart City 360°: First EAI International Summit, Smart City 360°, Bratislava, Slovakia and Toronto, Canada, October 13-16, 2015 : revised selected papers*, ed. A. Leon-Garcia, R. Lenort, D. Holman, D. Staš, V. Krutilova, P. Wicher, D. Cagánová, D. Špírková, et al., 592–597. [Cham]: Springer International Publishing.
- Barrington-Leigh, C., and A. Millard-Ball. 2019. Correction: The world's user-generated road map is more than 80% complete. *PloS one* 14. doi: 10.1371/journal.pone.0224742
- Berrios Villalba, A. 2020. How to Speed Up Digitization in the Railway. *IEEE Electrification Magazine* 8: 75–76.
- Campello, R.J.G.B., D. Moulavi, and J. Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*, ed. D. Hutchison, T. Kanade, J. Kittler, J.M. Kleinberg, F. Mattern, J.C. Mitchell, M. Naor, O. Nierstrasz, et al., 160–172. Berlin, Heidelberg: Springer.
- Cardone, G., A. Cirri, A. Corradi, L. Foschini, R. Ianniello, and R. Montanari. 2014. Crowdsensing in Urban Areas for City-Scale Mass Gathering Management: Geofencing and Activity Recognition. *IEEE Sensors Journal* 14: 4185–4195. doi: 10.1109/JSEN.2014.2344023
- Chang, K.-T. 2016. *Introduction to geographic information systems*. New York: McGraw-Hill Education.
- Cheng, G., Y. Guo, Y. Chen, and Y. Qin. 2019. Designating City-Wide Collaborative Geofence Sites for Renting and Returning Dock-Less Shared Bikes. *IEEE Access* 7. doi: 10.1109/ACCESS.2019.2903521
- Chen, M., M. Milenkovic, and M. Prosen. 2018. Smart Supply Chain Oriented Rail Freight Services: Smart-Rail recommendations for Shift2Rail. Retrieved 01.07.21, from <https://cordis.europa.eu/project/id/636071/results>.
- Coral, R., F. Esposito, and J. Weinstock. 2020. Don't Go There: A Zero-Permission Geofencing App to Alleviate Gambling Disorders. In *2020 IEEE 17th Annual Consumer*, 1–6.
- Corbetta, M., C. Sbarufatti, A. Manes, and M. Giglio. 2015. Real-Time Prognosis of Crack Growth Evolution Using Sequential Monte Carlo Methods and Statistical Model Parameters. *IEEE Transactions on Reliability* 64: 736–753. doi: 10.1109/TR.2014.2366759
- Cosulich, G., A. Derito, M. Giannettoni, and S. Savio. 2006. Results of the evaluation of F-MAN – An innovative solution for the management of railway cargo fleets. *IFAC Proceedings Volumes* 39: 331–336. doi: 10.3182/20060829-3-NL-2908.00058
- Daamen, W., R.M.P. Goverde, and I.A. Hansen. 2008. Non-Discriminatory Automatic Registration of Knock-On Train Delays. *Networks and Spatial Economics* 9: 47–61.
- Das, M., R. Cui, D.R. Campbell, G. Agrawal, and R. Ramnath. 2015. Towards methods for systematic research on big data. In *2015 IEEE International Conference on Big Data (Big Data)*, 2072–2081.
- DB Cargo AG. 2021a. Experiences and internal communications with DB Cargo employees.
- DB Cargo AG. 2021b. Ihr Logistikpartner. Retrieved 18.03.21, from <https://logistikpartner.dbcargo.com/>.
- Dong, X., M. Zhang, S. Zhang, X. Shen, and B. Hu. 2019. The analysis of urban taxi operation efficiency based on GPS trajectory big data. *Physica A: Statistical Mechanics and its Applications* 528. doi: 10.1016/j.physa.2019.121456
- European Commission. 2001. *Directive 2001/16/EC of the European Parliament and of the Council of 19 March 2001 on the interoperability of the trans-European conventional rail system*.
- European Commission. 2004. *Directive 2004/49/EC of the European Parliament and of the Council of 29 April 2004 on safety on the Community's railways and amending Council Directive 95/18/EC on the licensing of railway undertakings and Directive 2001/14/EC on the allocation of railway infrastructure capacity and the levying of charges for the use of railway infrastructure and safety certification (Railway Safety Directive)*.
- European Commission. 2011. White Paper: Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system. Retrieved 18.03.21, from <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52011DC0144&from=EN>.
- European Commission. 2021. Infrastructure - TEN-T - Connecting Europe: Corridors. Retrieved 10.08.21, from [https://ec.europa.eu/transport/themes/infrastructure/ten-t-guidelines/corridors\\_en](https://ec.europa.eu/transport/themes/infrastructure/ten-t-guidelines/corridors_en).
- European Space Agency. 2011a. BeiDou Performances. Retrieved 17.08.21, from [https://gssc.esa.int/navipedia/index.php/BeiDou\\_Performances#:~:text=The%20BeiDou%20System%20has%20been,within%200.2%20meters%20per%20second](https://gssc.esa.int/navipedia/index.php/BeiDou_Performances#:~:text=The%20BeiDou%20System%20has%20been,within%200.2%20meters%20per%20second).

- European Space Agency. 2011b. Galileo Performances. Retrieved 17.08.21, from [https://gssc.esa.int/navipedia/index.php/Galileo\\_Performances](https://gssc.esa.int/navipedia/index.php/Galileo_Performances).
- Eurostat. 2020. Freight transport statistics. Retrieved 21.01.21, from [https://ec.europa.eu/eurostat/statistics-explained/index.php/Freight\\_transport\\_statistics#Rail\\_transport](https://ec.europa.eu/eurostat/statistics-explained/index.php/Freight_transport_statistics#Rail_transport).
- Gaur, C. 2020. Top 6 Big Data Challenges and Solutions to Overcome. Retrieved 03.08.21, from <https://www.xenonstack.com/insights/big-data-challenges>.
- GCU. 2020. Signatories – GCU. Retrieved 17.03.21, from <https://gcubureau.org/signatories-2/>.
- Ghofrani, F., Q. He, R.M. Goverde, and X. Liu. 2018. Recent applications of big data analytics in railway transportation systems: A survey. *Transportation Research Part C: Emerging Technologies* 90: 226–246. doi: 10.1016/j.trc.2018.03.010
- Goverde, R., and I. Hansen. 2000. TNV-Prepare: Analysis of Dutch railway operations based on train detection data. *Computers in Railways VII* 50: 780–788.
- Grand View Research. 2021. Railway Telematics Market Size, Share & Trends Analysis Report By Solution (Fleet Management, Automatic Stock Control, Remote Data Access, Railcar Tracking & Tracing), By Railcar, By Component Type, And Segment Forecasts, 2021 - 2028. Retrieved 03.10.21, from <https://www.grandviewresearch.com/industry-analysis/railway-telematics-market-report>.
- Guha, S., R. Rastogi, and K. Shim. 1998. *CURE: an efficient clustering algorithm for large databases*.
- Gülagiz, F.K., and S. Sahin. 2017. *Comparison of Hierarchical and Non-Hierarchical Clustering Algorithms*. Dubai: Dorma Trading, Est. Publishing Manager.
- Hartigan, J.A. 1975. *Clustering algorithms*. New York: Wiley.
- International Transport Forum. 2015. The carbon footprint of global trade: Tackling emissions from international freight transport. Retrieved 13.03.21, from <https://www.itf-oecd.org/sites/default/files/docs/cop-pdf-06.pdf>.
- Ishioka, T. 2006. *An expansion of x-means: Progressive iteration of k-means and merging of the clusters*. Japanese Society of Computational Statistics.
- Ishwarappa, K., and J. Anuradha. 2015. A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. *Procedia Computer Science* 48: 319–324. doi: 10.1016/j.procs.2015.04.188
- Islam, D.M.Z., S. Ricci, and B.-L. Nelldal. 2016. How to make modal shift from road to rail possible in the European transport market, as aspired to in the EU Transport White Paper 2011. *European Transport Research Review* 8. doi: 10.1007/s12544-016-0204-x
- Jagadish, H.V. 2015. Big Data and Science: Myths and Reality. *Big Data Research* 2: 49–52. doi: 10.1016/j.bdr.2015.01.005
- Jin, X., B.W. Wah, X. Cheng, and Y. Wang. 2015. Significance and Challenges of Big Data Research. *Big Data Research* 2: 59–64. doi: 10.1016/j.bdr.2015.01.006
- Kitchin, R. 2014. *Big Data, new epistemologies and paradigm shifts*. London, England: SAGE Publications.
- Kresse, W., and D.M. Danko. 2012. *Springer handbook of geographic information*. Berlin, New York: Springer.
- Liu, S., Le Yin, W.K. Ho, K.V. Ling, and S. Schiavon. 2017. A tracking cooling fan using geofence and camera-based indoor localization. *Building and Environment* 114: 36–44. doi: 10.1016/j.buildenv.2016.11.047
- Li, Z., and Q. He. 2015. Prediction of Railcar Remaining Useful Life by Multiple Data Source Fusion. *IEEE Transactions on Intelligent Transportation Systems* 16: 2226–2235. doi: 10.1109/TITS.2015.2400424
- Maass, W., J. Parsons, S. Puro, V.C. Storey, and C. Woo. 2018. Data-Driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research. *Journal of the Association for Information Systems*: 1253–1273. doi: 10.17705/1jais.00526
- Ma, E.W., and T.W. Chow. 2004. A new shifting grid clustering algorithm. *Pattern Recognition* 37: 503–514. doi: 10.1016/j.patcog.2003.08.014
- Maimon, O.Z., and L. Rokach. 2005. *Data mining and knowledge discovery handbook*. Ramat-Aviv, Great Britain: Springer.
- Malle, J.-P. 2013. Big Data: Farewell to Cartesian Thinking? Retrieved 03.08.21, from <http://www.paristechreview.com/2013/03/15/big-data-cartesian-thinking/>.
- Marinov, M. 2018. *Sustainable rail transport: Proceedings of RailNewcastle Talks 2016*. Cham: Springer.
- Mazzocchi, F. 2015. Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO reports* 16: 1250–1255. doi: 10.15252/embr.201541001
- Medeossi, G., G. Longo, and S. de Fabris. 2011. A method for using stochastic blocking times to improve timetable planning. *Journal of Rail Transport Planning & Management* 1: 1–13. doi: 10.1016/j.jrtpm.2011.07.001
- Merry, K., and P. Bettinger. 2019. Smartphone GPS accuracy study in an urban environment. *PloS one* 14: e0219890. doi: 10.1371/journal.pone.0219890
- Miller, H. 2007. Place-Based versus People-Based Geographic Information Science. *Geography Compass* 1: 503–535. doi: 10.1111/j.1749-8198.2007.00025.x
- Miller, H.J. 2010. The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science* 50: 181–201. doi: 10.1111/j.1467-9787.2009.00641.x
- Miller, H.J., and M.F. Goodchild. 2015. Data-driven geography. *GeoJournal* 80: 449–461. doi: 10.1007/s10708-014-9602-6

- Millward, R. 2005. *Private and public enterprise in Europe: Energy, telecommunications and transport, 1830-1990*. Cambridge UK, New York: Cambridge University Press.
- Mirzabeiki, V., J. Holmström, and P. Sjöholm. 2012. Aligning organisational interests in designing rail-wagon tracking. *Operations Management Research* 5: 101–115. doi: 10.1007/s12063-012-0072-z
- NASA. 2021. A Landsat Timeline. Retrieved 03.08.21, from <https://landsat.gsfc.nasa.gov/about/landsat-timeline>.
- NetRail AB. 2021. Fordons- & vagnservice. Retrieved 30.07.21, from <https://netrail.se/fordonservice/>.
- Nunez, A., J. Hendriks, Z. Li, B. de Schutter, and R. Dollevoet. 2014. Facilitating maintenance decisions on the Dutch railways using big data: The ABA case study. In *IEEE International Conference on Big Data*, 48–53. IEEE.
- Odlyzko, A. 2010. Collective Hallucinations and Inefficient Markets: The British Railway Mania of the 1840s. *SSRN Electronic Journal*. doi: 10.2139/ssrn.1537338
- Oxford Dictionary of English*. 2010. Oxford University Press.
- Papaelias, M., A. Amini, Z. Huang, P. Valley, D.C. Dias, and S. Kerkyras. 2016. Online condition monitoring of rolling stock wheels and axle bearings. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit* 230: 709–723. doi: 10.1177/0954409714559758
- Pérez-Fernández, O., and J.C. García-Palomares. 2021. Parking Places to Moped-Style Scooter Sharing Services Using GIS Location-Allocation Models and GPS Data. *ISPRS International Journal of Geo-Information* 10: 230. doi: 10.3390/ijgi10040230
- Python Software Foundation. 2021. DiffliB.py. Retrieved 08.08.21, from <https://github.com/python/cpython/blob/master/Lib/diffliB.py>.
- Rabiei-Dastjerdi, H., G. McArdle, and A. Ballatore. 2020. Urban Consumption Patterns: OpenStreetMap Quality for Social Science Research. In *Proceedings of the 6th International Conference on Geographical Information Systems Theory, Applications and Management*, 278–285. SCITEPRESS - Science and Technology Publications.
- Rail Cargo Group. 2019. How does cross-border transport work? Retrieved 22.03.21, from <https://blog.railcargo.com/en/artikel/eisenbahn-einfach-erklart-interoperabilitaet/eisenbahn-einfach-erklart-grenzueberschreitender-verkehr>.
- Railnova. 2021. Predictive maintenance for railway fleets. Retrieved 03.10.21, from <https://www.railnova.eu/>.
- Railway Technology. 2018. The world's biggest railway operators in 2018. Retrieved 18.03.21, from <https://www.railway-technology.com/features/worlds-biggest-railway-operators-2018/>.
- Ramirez, R.C., I. Moya, I. Puy, U. Alvarado, I. Adin, and J. Mendizabal. 2020. Freight Telematics Systems: An Intelligent Wagon. In *Communication Technologies for Vehicles*, ed. F. Krief, H. Aniss, L. Mendiboure, S. Chaumette, and M. Berbineau, 157–165. Cham: Springer International Publishing.
- Ratcliff, J.W., and D.M. Metzener. 1988. Gestalt: An introduction to the Ratcliff/Obershelp pattern matching algorithm. *Dr. Dobbs Journal* 7.
- Research and Markets. 2021. Railway Telematics Market by Solution (Fleet Management, Automatic Stock Control, Shock Detection, Reefer Wagon Management, ETA), Railcar (Hoppers, Tank Cars, Well Cars, Boxcars, Reefer Cars), Component & Region - Global Forecast to 2026. Retrieved 03.10.21, from <https://www.marketsandmarkets.com/Market-Reports/railway-telematics-market-182239471.html>.
- Reclus, F., and K. Drouard. 2009. Geofencing for fleet & freight management. In *9th International Conference on Intelligent Transport Systems Telecommunications*, 353–356.
- Roth, R., and G. Dinobl. 2008. *Across the Borders: Financing the World's Railways in the Nineteenth and Twentieth Centuries*. Burlington: Ashgate Publishing Limited.
- Roth, R., and C. Divall. 2015. *From Rail to Road and Back Again?: A Century of Transport Competition and Interdependency*. London: Taylor and Francis, 446 pp.
- Sammouri, W., E. Côme, L. Oukhellou, and P. Aknin. 2013. Mining Floating Train Data Sequences for Temporal Association Rules within a Predictive Maintenance Framework. In *Advances in Data Mining. Applications and Theoretical Aspects: 13th Industrial Conference, ICDM 2013, New York, NY, USA, July 16-21, 2013. Proceedings*, ed. P. Perner, 112–126. Berlin, Heidelberg: Springer.
- Schiller, T., M. Maier, and M. Büchle. 2017. Global Truck Study 2016: The truck industry in transition. Retrieved 10.08.21, from <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/manufacturing/us-manufacturing-global-truck-study-the-truck-industry-in-transition.pdf>.
- Scikit-Learn Python Library. 2021. Comparing different clustering algorithms on toy datasets. Retrieved 26.09.21, from [https://scikit-learn.org/stable/auto\\_examples/cluster/plot\\_cluster\\_comparison.html](https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html).
- Shmueli, and Koppius. 2011. Predictive Analytics in Information Systems Research. *MIS Quarterly* 35: 553. doi: 10.2307/23042796
- Sieferle, R.P. 2008. *Transportgeschichte*. Berlin: LIT Verlag, 292 pp.
- Šimeková, Ž., M. Balog, and A. Rosová. 2013. The use of the RFID in rail freight transport in the world as one of the new technologies of identification and communication. *Acta Montanistica Slovaca* 18: 26–32.
- Sims, R., R. Schaeffer, F. Creutzig, X. Cruz-Núñez, M. D'Agosto, D. Dimitrou, M. Figueroa Meza, L. Fulton, et al. 2014. Transport. In *Climate change 2014 - Mitigation of climate change: Working Group III*

- contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. O. Edenhofer, R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, et al. Cambridge, UK, New York, USA: Cambridge University Press.
- Strasser, B.J. 2012. Data-driven sciences: From wonder cabinets to electronic databases. *Studies in history and philosophy of biological and biomedical sciences* 43: 85–87. doi: 10.1016/j.shpsc.2011.10.009
- Suchanek, M., ed. 2017. *Sustainable Transport Development, Innovation and Technology: Proceedings of the 2016 TranSopot Conference*. Cham: Springer International Publishing.
- Sui, D., and M. Goodchild. 2011. The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science* 25: 1737–1748. doi: 10.1080/13658816.2011.604636
- SweMaint AB. 2021. From South to North: 12 wagon workshops, 30 wheel supply locations, 40 mobile units. Retrieved 02.08.21, from <https://www.swemaint.com/en/workshops>.
- U.S. Space Force. 2021. GPS Accuracy: How accurate is GPS? Retrieved 17.08.21, from <https://www.gps.gov/systems/gps/performance/accuracy/>.
- Umweltbundesamt. 2019. Emissionsdaten: Emissionen im Güterverkehr - Tabelle. Retrieved 13.03.21, from <https://www.umweltbundesamt.de/themen/verkehr-laerm/emissionsdaten#tabelle>.
- USGS. 2021. What is a geographic information system (GIS)? Retrieved 05.08.21, from [https://www.usgs.gov/faqs/what-a-geographic-information-system-gis?qt-news\\_science\\_products=0#qt-news\\_science\\_products](https://www.usgs.gov/faqs/what-a-geographic-information-system-gis?qt-news_science_products=0#qt-news_science_products).
- Welles, B.K., and E.J. Hershey. 1997. Use of mutter mode in asset tracking for gathering data from cargo sensors, US 5,686,888.
- Yin, J., and W. Zhao. 2016. Fault diagnosis network design for vehicle on-board equipments of high-speed railway: A deep learning approach. *Engineering Applications of Artificial Intelligence* 56: 250–259. doi: 10.1016/j.engappai.2016.10.002
- Zarembski, A.M., D. Einbinder, and N. Attoh-Okine. 2016. Using multiple adaptive regression to address the impact of track geometry on development of rail defects. *Construction and Building Materials* 127: 546–555. doi: 10.1016/j.conbuildmat.2016.10.012
- Zhang, T., R. Ramakrishnan, and M. Livny. 1996. *BIRCH: an efficient data clustering method for very large databases*.
- Zhang, Y., and Z. Mi. 2018. Environmental benefits of bike sharing: A big data-based analysis. *Applied Energy* 220: 296–301. doi: 10.1016/j.apenergy.2018.03.101

## Appendix

Appendix A: Number of maintenance facilities in Europe included in VPI and DB Cargo lists per country

Appendix B: ArcGIS Pro models used to create infrastructure geofences

Appendix C: ArcGIS Pro models used to create telematics geofences.

Appendix D: ArcGIS Pro models used to combine the infrastructure and telematics geofences.

## Appendix A

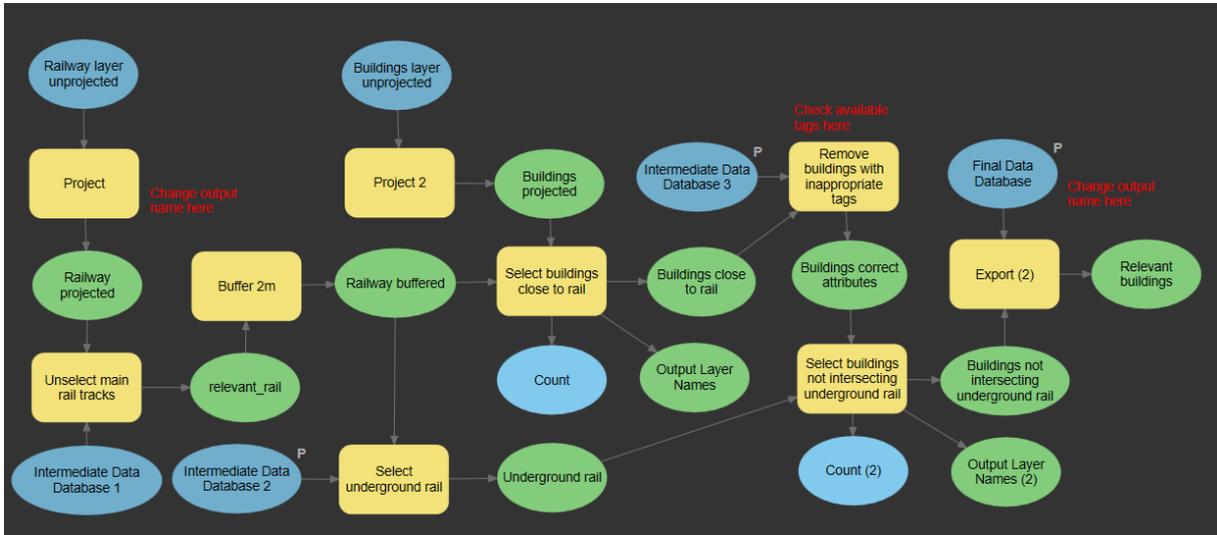
Number of maintenance facilities in Europe included in VPI and DB Cargo lists per country and in total. As duplicate cities were removed, the number of entries in the combined dataset can be lower than the sum of the two lists.

Country	Number of maintenance facilities		
	VPI	DB Cargo	Combined dataset
Austria	3	5	7
Belgium	1	7	7
Bulgaria	6	2	5
Croatia	0	1	1
Czech Republic	3	6	8
France	13	26	39
Germany	60	85	118
Hungary	3	5	7
Italy	5	12	16
Luxemburg	0	1	1
Netherlands	3	7	7
Norway	3	0	3
Poland	12	12	22
Romania	8	3	9
Serbia	1	2	2
Slovakia	0	2	2
Slovenia	0	2	2
Spain	16	0	15
Sweden	4	11	12
Switzerland	1	10	10
United Kingdom	21	0	21
<b>Total</b>	<b>163</b>	<b>199</b>	<b>314</b>

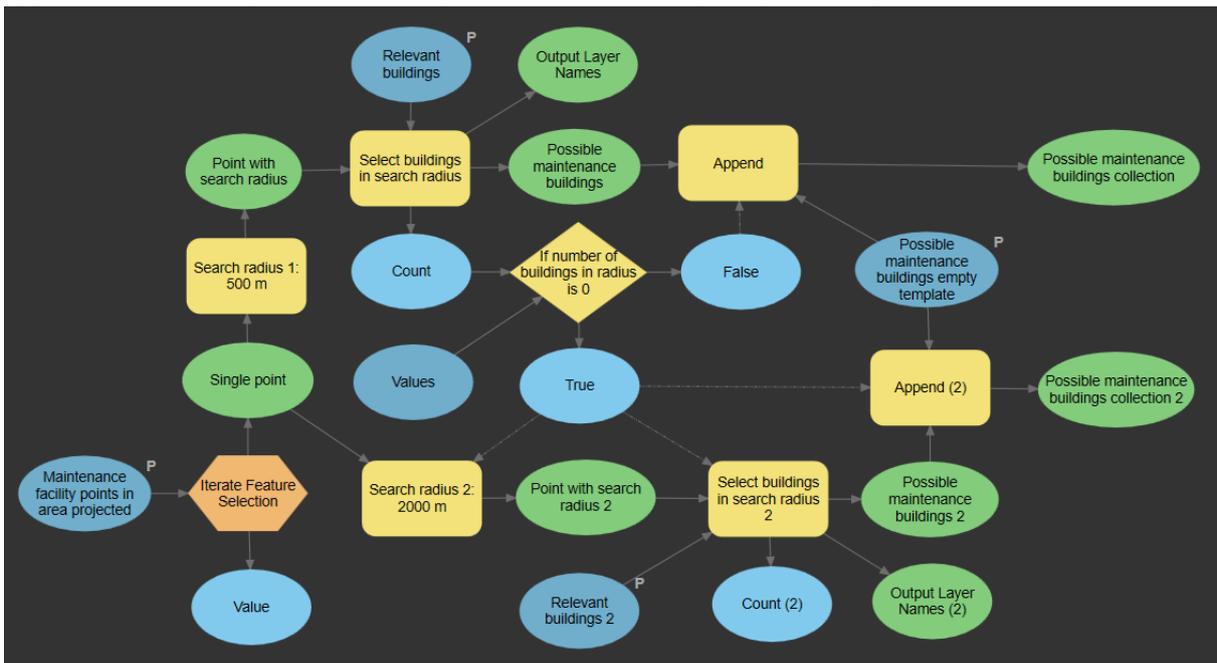
## Appendix B

ArcGIS Pro models used to create infrastructure geofences.

Appendix B.1: Preprocessing of OSM rail and buildings data.



Appendix B.2: Selecting relevant buildings close to maintenance facility address points.





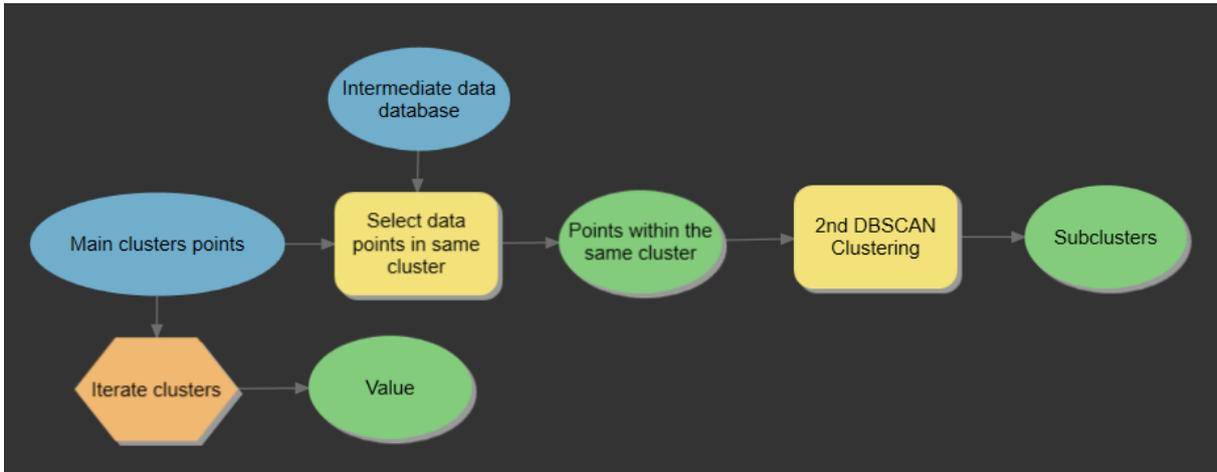
## Appendix C

ArcGIS Pro models used to create telematics geofences.

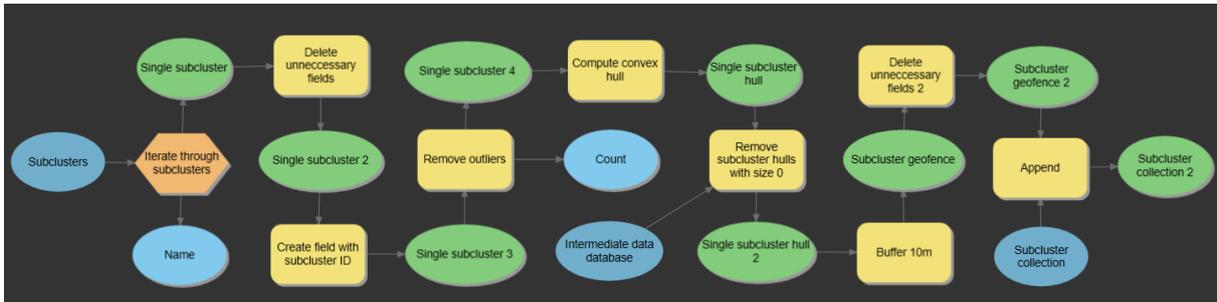
*Appendix C.1: First DBSCAN clustering of GNSS points of damaged wagons. Search radius = 100 m, minimum number of points per cluster = 20.*



*Appendix C.2: Second DBSCAN clustering applied to each cluster of the first clustering round. Search radius = 50 m, minimum number of points per cluster = 10.*



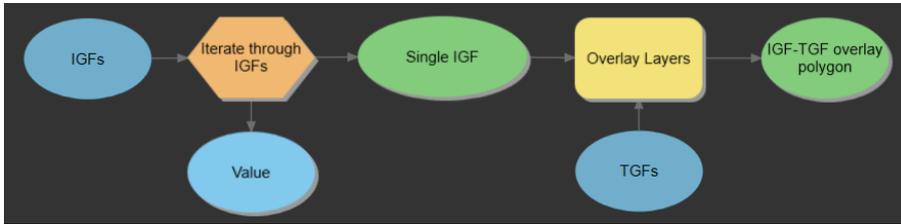
*Appendix C.3: Removing outliers and small clusters and computing the convex hull for the subclusters to create the GNSS geofences.*



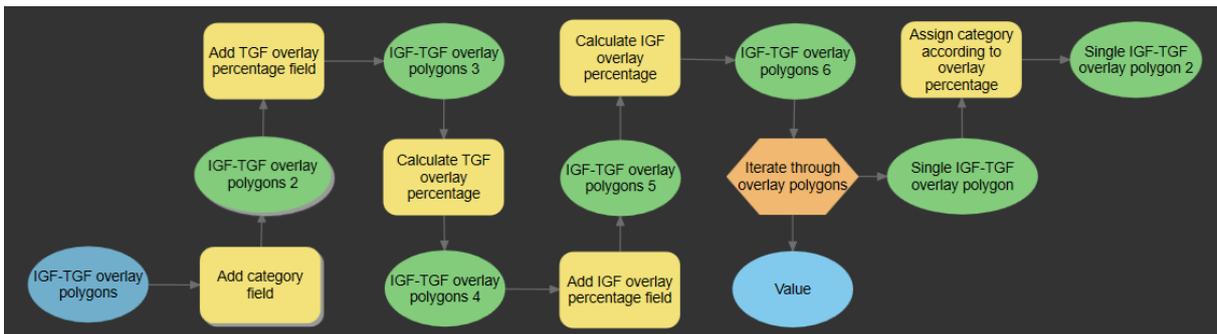
## Appendix D

ArcGIS Pro models used to combine the infrastructure and telematics geofences.

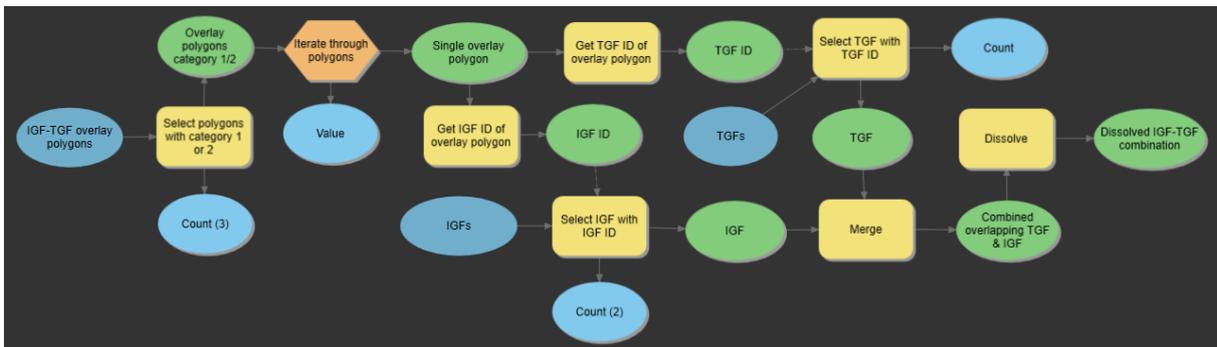
Appendix D.1: Overlaying infrastructure and telematics geofences.



Appendix D.2: Calculating the overlay area percentage of each overlay.



Appendix D.3: Merging infrastructure and telematics geofences if they overlap a certain amount.



Appendix D.4: Deleting infrastructure geofences if other overlapping geofences or telematics geofences are close by.

