Predicting infrastructural failures with vehicle data

LUND UNIVERSITY
Campus Helsingborg

LTH School of Engineering at Campus Helsingborg
Department of civil engineering

Bachelor thesis:
Cajsa Åkerlund
Linus Rydstedt

LTH School of Engineering
Lund University
Box 882
SE-251 08 Helsingborg
Sweden

LTH Ingenjörshögskolan vid Campus Helsingborg
Lunds universitet
Box 882
251 08 Helsingborg

Printed in Sweden
Media-Tryck
Biblioteksdirektionen
Lunds universitet
Lund 2020

Abstract

In this thesis, we have investigated if infrastructural failures can be predicted using event codes generated by the train computer. For this purpose, historical data with known infrastructural failures have been examined together with event codes generated by the vehicle. We have then tried to find a pattern using relative risk. A confidence interval was calculated for the relative risk to see if the connection was clear enough. After the statistical analysis the event codes have been further investigated to see if the connection is possible.

The infrastructural failures that have been investigated are *Cant irregularities*, *Incorrect position of the contact wire* and *Damage and wear on the contact wire*. We have not found any connection between *Cant irregularities* and the event codes. However, we have found that both *Incorrect position of the contact wire* and *Damage and wear on the contact wire* are associated with certain event codes. These event codes have a correlation, but it is not clear if they can predict an infrastructural malfunction or if they indicate a behavior that may be damaging to the infrastructure. These event codes need to be further investigated to see if they can be used to predict an infrastructural failure. We have examined what the information process looks like today and concluded that an automatic alert should be sent to the dispatcher from Alstom if the event codes can predict an infrastructural failure.

Keywords
Predicting, Infrastructure failures, Railway, Trains, Train-Track

Sammanfattning

I examenarbetet undersöks det om man kan upptäcka fel på infrastrukturen med hjälp av eventkoder genererade från fordonsdatorn. För att kunna dra en slutsats om man kan förutspå ett infrastrukturfel med hjälp av fordonsdata har historiska data med kända fel från infrastrukturen undersökts tillsammans med eventkoder som genererats av fordonen. Vi har sedan försökt att hitta ett samband med hjälp av relativ risk. En kontroll om sambandet är relevant eller inte har gjorts genom att använda konfidensintervall. Därefter har eventkodernas betydelse undersökts för att se om sambandet är rimligt.

De infrastrukturfel som har undersökts är *Skevningsfel*, *Felaktig position av kontaktledning* samt S*kador och slitage på kontaktledning*. Vi har inte funnit något samband mellan *Skevningsfel* och event koder. Däremot har vi hittat ett samband mellan eventkoderna och *Felaktig positionering av kontaktledningen* samt *Skador och slitage på kontaktledningen*. Vi har funnit ett samband mellan eventkoderna och infrastrukturfelen men det är inte fastställt om eventkoderna kan användas för att förutspå ett infrastrukturfel eller om de indikerar på ett felaktigt förarbeteende som eventuellt kan skada infrastrukturen. De här eventkoderna behöver undersökas ytterligare för att fastställa att de kan användas för att förutspå infrastrukturproblem. Vi har även undersökt hur informationsprocessen ser ut idag och kommit fram till att ett automatiskt utskick till tågtrafikledaren från Alstom rekommenderas om eventkoderna kan upptäcka ett infrastrukturfel.

Nyckelord
Förutspå, Infrastrukturfel, Järnväg, Tåg, Fordon-Bana

Preface

This thesis is the final part of our Bachelor of Science in Engineering, Civil Engineering, at Faculty of Engineering LTH at Lund University. The thesis has been in a collaboration with Alstom Transport during the spring of 2020.

First of all, we would like to take the opportunity to thank all those who have been involved in this thesis, Daniel Petersson at Alstom for helping us with the event codes and Anna Olsson at Arriva for taking time to explain how information is distributed between the companies. Furthermore, we highly appreciate all assistance in the data collection process, without the data this thesis would not have been possible. Thanks to Leif Nilsson, Peter Isaksson and Mats Wilhelmsson at Trafikverket for helping us collect infrastructural data and involved personnel at Skånetrafiken for approving that we could use the vehicle data. Hopefully, this thesis will be useful for readers and future research about similar topics or related field works.

Last but not least, we want to specially emphasize our thanks to our supervisor Carl-William Palmqvist, Postdoctoral fellow at LTH and Jon Hankers, RAM Engineer at Alstom for their guidance, support and help in order to create this thesis.

Table of contents

1. Introduction and background

Train traffic on the Swedish railway network has increased over the recent years (Gummeson, 2019). Track maintenance has not followed at the same rate which has created disruptions in the railway system (Honauer & Ödeen, 2019). To manage railway traffic efficiently, it is crucial to have continuous maintenance on the infrastructure and the trains. To be able to detect a potential failure, both the infrastructure and trains are equipped with sensors. The data generated by the vehicle sensors are used for planning maintenance on the vehicles. This bachelor thesis investigates whether it is also possible to predict infrastructural problems using data from the vehicles.

This thesis is a continuation of a previous sprint in an organization called Together for Trains on Time (TTT) (Finn, 2019), developed by Järnvägsbranchens samverkansforum (JBS) (collaboration forum for the railway industry). The purpose of TTT is to encourage railway companies to cooperate and create a more reliable railway system. In the sprint, vehicle data from Alstom was used to find relations and trends between vehicle data and noted disruptions in the infrastructure, particularly focusing on problems occurring on the catenary. Vehicle data from Alstom has not previously been used to predict any other types of infrastructure problems in Sweden, but it is known that infrastructural problems can generate damage on a vehicle, for example, badly ground rails can generate unnecessary wear (Lundmark, 2007). If infrastructural problems are detected at an early stage, these problems can be fixed before traffic is affected. This also enables systematic maintenance to minimize the damage, and ultimately impacts the long-term planning for the Swedish railway.

A collaboration project between Trafikverket, Region Stockholm and MTR (2019) showed that vehicles can detect problems occurring at the contact wire if they are carrying measuring equipment. One recent study by De Rosa et. al. (2020), shows that machine learning algorithms can be used to identify lateral track irregularities. Previous studies tend to investigate vertical issues. Karis (2017) investigated the correlation of track irregularities and vehicle responses based on simulations, using data from simulations from two separate projects to evaluate track-vehicle interactions into three parts and these are analyzed in terms of correlation. Systems to measure track geometry using in-service vehicles have been developed and more common over the resent years (Weston, Roberts, Yeo, & Stewart, 2015). For example Alstom has developed in-service measuring equipment, but nothing has been installed on Swedish vehicles but there are active systems in other countries (Alstom, 2016). The information

measured and generated from the vehicles enables the opportunity for the maintenance companies to plan their work. By using bogie-mounted sensors Weston, et.al., (2007) mentioned in recent studies a technique to assess wavelength from track irregularities.

1.1. Aim

The aim is to use statistical methods to find a link between abnormal behavior in the train computer at locations where there is a known infrastructural failure. Furthermore, to investigate if the generated event codes can predict the infrastructural problem in the future. The aim is also to suggest future development ideas and suggest which actors should have access to the eventual warning system.

1.2. Problem

- Is it possible to sort the vehicle- and inspections data in a way that it can be analyzed?
- Is the pattern in the sorted data clear enough to be able to predict infrastructural failures in the future?
- Is it possible to get a more reliable reading by adding data from other sources, for example track information data?
- What/Which actor(s) should get access to the warning system that the algorithms will trigger, and where in the information flow should the system send a warning?

1.3. Limitations

The data and analysis are limited to the commuter trains (Pågatåg) in Skåne, from 2017-01-01 to 2019-06-30. Infrastructure inspection data are available from 2017-01-02 to 2019-12-30.

The thesis only considers failures categorized as *Cant irregularities (twists)*, *Incorrect position of the contact wire* and *Damage and wear* on *the contact wire.*

In this thesis the event codes are only investigated to see which event codes that did not have a connection to any of the infrastructural failures.

2. Technical background

To get a reliable railway system, there are many sub-systems that must cooperate. This chapter contains facts about the trains, followed up by an explanation about the information flow between different actors. Finally, it includes information about the considered infrastructural problems, and how inspection data is generated.

2.1. Commuter trains in Skåne and vehicle event codes

Pågatåget, Skåne's commuter train, is part of the Coradia Nordic-series. There are 99 Pågatåg trains, 10 of which are on loan to Västtrafik. The trains are equipped with sensors that continuously send data to Alstom's servers. This data is stored and used as a basis when planning maintenance and service work. These trains run on eight different routes on the railway network in Skåne (Skånetrafiken, 2020). The trains run in different circulations each day, each circulation is about 1 000 km and there are over 60 circulations per day (Olsson, 2020). They operate on the following lines: 'Södra stambanan', 'Skånebanan', 'Västkustbanan', 'Lommabanan', 'Rååbanan', 'Ystadbanan', 'Trelleborgsbanan', 'Österlenbanan', 'Blekingebanan', 'Markarydbanan', 'Citytunneln' and 'Kontinentalbanan'. The two lines with the most traffic, are Södra stambanan and Västkustbanan, shown in the picture below. There are 260 000 train services using Södra stambanan every year, which is 27% of all train movements in Sweden (Corshammar, 2006). Therefore, predicting an infrastructural problem here could have a major positive impact on Swedish railway.

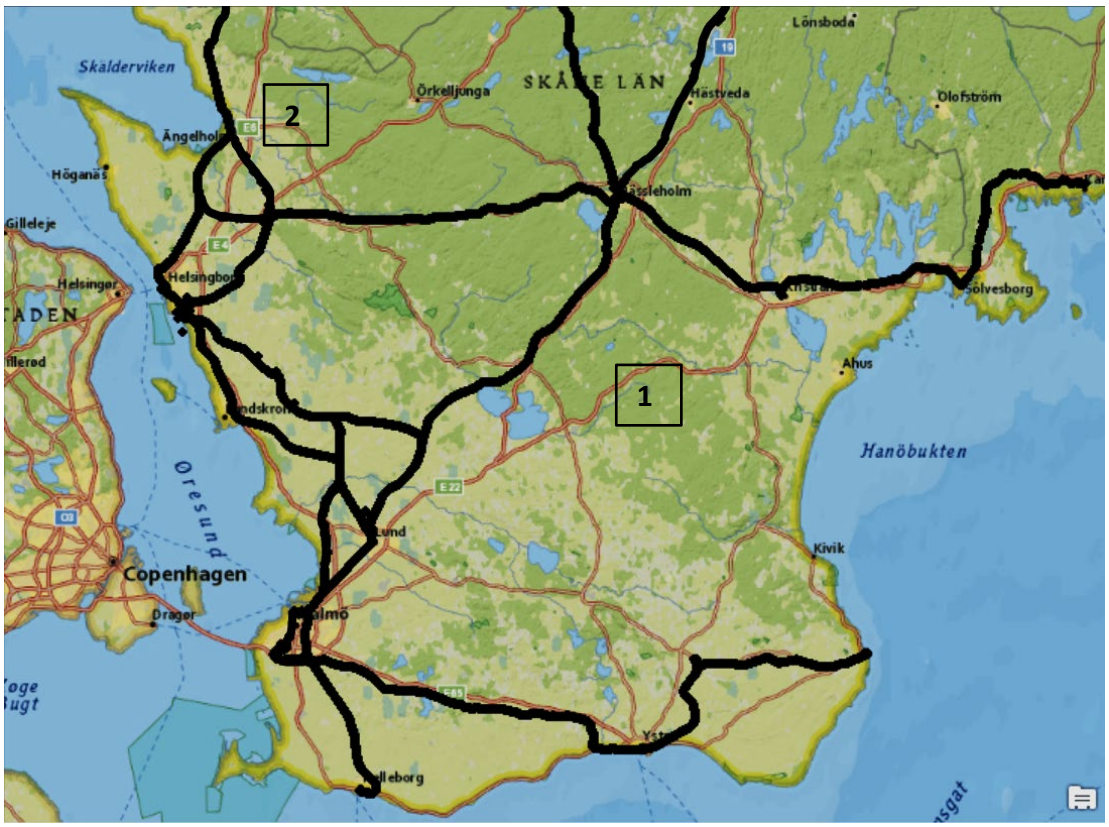

Figure 1 Södra stambanan (1), Västkustbanan (2).

The data used from the vehicles consists of generated event codes. There are 734 different types of event codes that have been generated by the vehicles during the investigated time. In this time period the event codes have been generated around four million times. The event codes vary in severity: Some need urgent attention, some are information messages, like if the vehicle is handled in a way that differs from normal behavior, and some are messages about something that needs manual attention, like low fresh water. The vehicle event codes are not mainly intended to detect a failure but to inform maintenance staff about the condition of the vehicle and what needs to be prioritized. Appendix A shows the event codes that figure in this thesis.

2.2. Different actors and information flow

Trafikverket owns and manages about 90% of the Swedish railway (Trafikverket, 2019). They are responsible for long-term infrastructural planning, construction of new railways, operation and maintenance of the infrastructure, managing traffic, and informing customers about the traffic situation. Trafikverket procures infrastructure maintenance contractors, who in turn are responsible for the quality level of the railway. The passenger transport executive owns or rents the vehicles for the services that they operate. The Transport executive assigns an operator who will have the major responsibility of the operation and maintenance of the vehicle. See Figure 2.

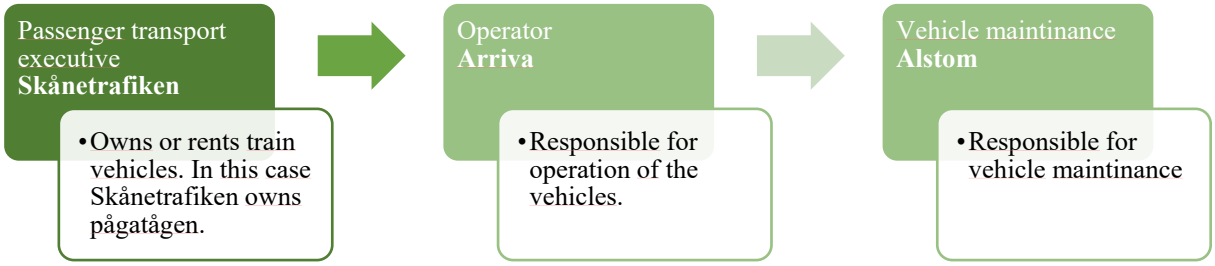

Figure 2 Description of the different actors.

Arriva is a public transport operator, which means that they have the responsibility for drivers, planning which lines the vehicles should run and planning maintenance for the vehicles. Apart from Skåne, Arriva also operates public transportation in the Stockholm region, Halmstad and Östergötland (Arriva, 2020). As illustrated in Figure 2, Arriva has procured Alstom as maintenance operator. Alstom is an international railway company that develops and sells railway infrastructure and vehicles and offers maintenance and support. In Sweden, Alstom maintains commuter trains in Skåne, Östergötland, and Västra Götaland.

In case of an infrastructure failure, information must be transmitted between the infrastructure manager, passenger transport executive, traffic operators and maintenance operators. Infrastructure failures can be detected by train drivers, sensors on the track, by manual inspections or inspection trains. The train driver is obligated to report any failure detected on the tracks to a train dispatcher. The train dispatcher takes necessary measures, for instance, to redirect the traffic and report the failure to a technician. An event is noted in Opal if the failure has caused a delay longer than 3 minutes. Opal dispatches an email to the train operators. The technician contacts the maintenance contractor who in turn is obligated to inspect the failure.

2.3. Infrastructural problems and data from the tracks

This thesis considers malfunctions that can be categorized as *Cant irregularities*, *Incorrect positioning of the contact wire* and *Damage and wear on the contact wire*. *Cant irregularities* are rail-tilt deviations from the design geometry and defined as the difference in cant between two cross-sections of the track, divided by the distance between these cross-sections (Andersson, Berg, & Stichel, 2014). If the tilt of the rail in two points with a specified distance, is too big or too small, there is an irregularity, se Figure 3. How much the rail is allowed to tilt depends on the speed limit of the railway line. If there is a minor irregularity, the train must run at reduced speed. *Cant irregularities* arise by forces in the rail, which are often created when the structure of the material is changed, for example by the weather. Track forces must be limited with respect to safety, maintenance, and passenger comfort. The main risk with *Cant irregularities* is that it can cause extensive damage to infrastructure, and in some extreme cases it can cause the trains to derail. As it affects the safety it is highly prioritized and should be remedied immediately.

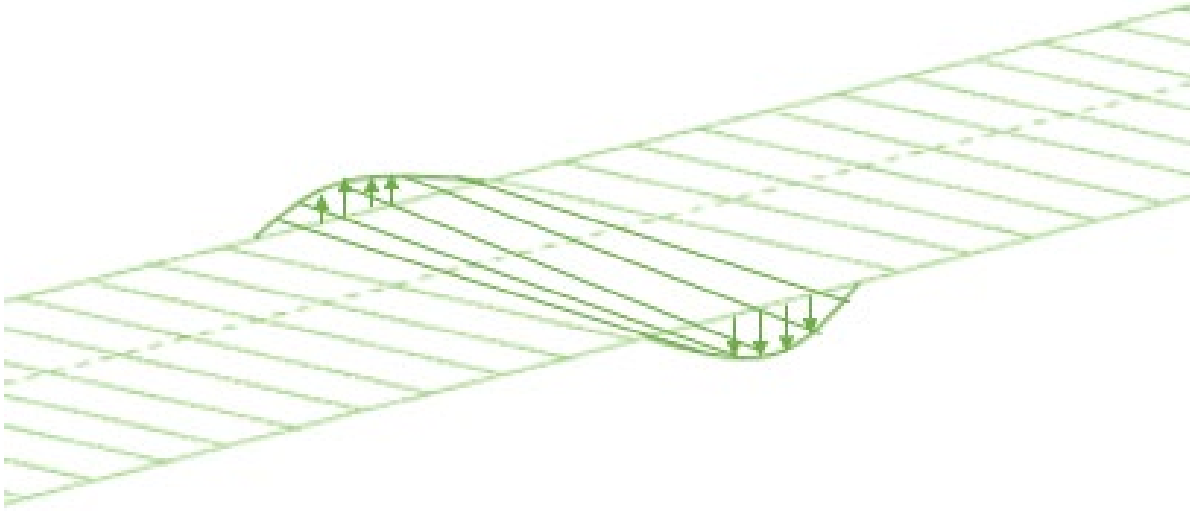

Figure 3 Illustration describing Cant irregularity.

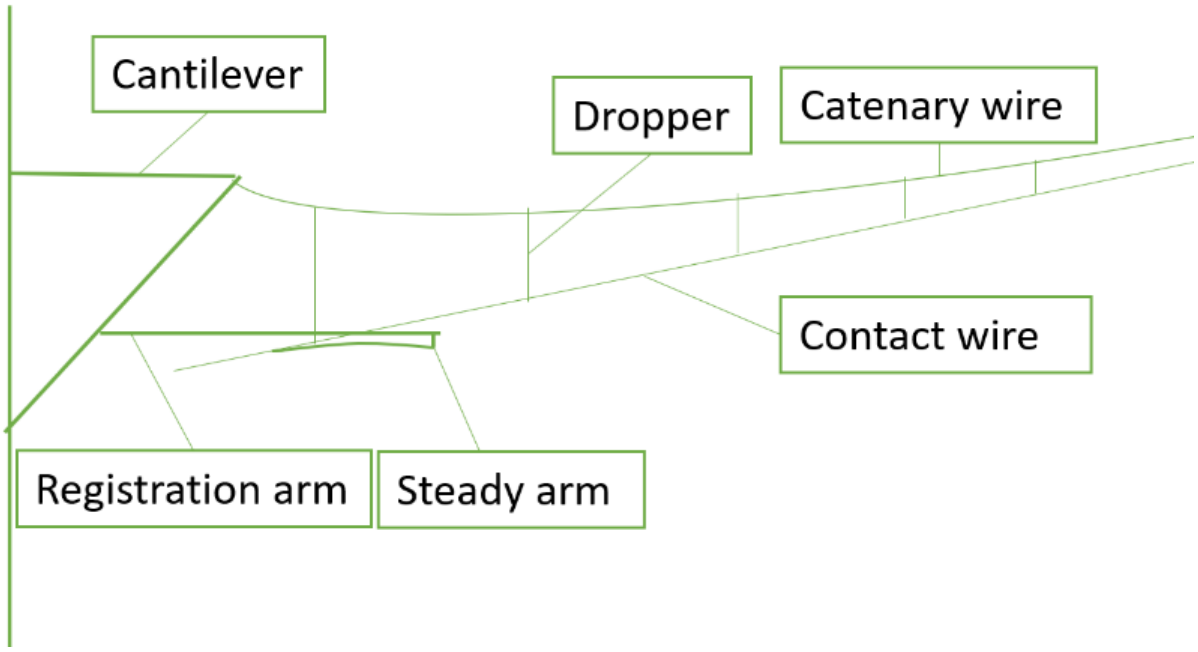

Figure 4 Illustration of different parts of the overhead line.

Incorrect position is when the position of the contact wire in relation to the center of the tracks is incorrect, malfunctions included in the category are mentioned in Appendix B. *Incorrect position of the contact wire* can be caused when the weight that tighten the catenary wire is too low, or if the steady arm is broken, se Figure 4. In some cases, the contact wire is positioned outside the carbon strips on the pantograph, this can lead to extensive wear on the pantographs and in some cases, incorrect positioning in horizontal length can lead to breakage of the contact wire. *Damage and wear on the contact wire* can be caused by several different reasons, the faults that are included in this category vary from spot wear to broken stands, all malfunctions that are included in the category are mentioned in Appendix B. For example, twists on the contact wire or vertical accelerations on the overhead line can produce spot wear. Wear on the overhead line can also be caused by bypassing vehicles if the vehicle´s carbon strips are too thin (Banverket, 2006). The worst-case scenario is that the contact wire can break if damage and wear is not detected in time.

3. Method

The method of this thesis can be divided into two different categories – "Data on infrastructural malfunctions" and "Calculating relative risk and confidence interval".

3.1. Data on infrastructure malfunctions

Data from the vehicles and inspections needs to be sorted before it can be analyzed. The data that was used when matching event codes from the vehicles to *Cant irregularities* was generated from Trafikverket's measuring vehicle. *Cant irregularities* were chosen because it was the most common problem if the data was sorted by the highest priority. In this thesis, all noted *Cant irregularities* are analyzed, not just the ones with high priority.

To analyze failures connected to the contact wire all data from Bessy during the investigated time limited to Skåne were used. Many datapoints connected to the same type of failure are needed to get a reliable result. A grouping of similar failures was made, which ended in two different groups, *Incorrect position of the contact wire* and *Damage and wear on the contact wire*.

For the categorizations, the first selection was made to only see the technical category *catenary* where the infrastructural failure had already been fixed. In Trafikverket's data, there is not one category that includes all problems connected to *Incorrect positioning of contact wire* or *Damage and wear on contact wire*, therefore a categorization has been made. The category *Incorrect position of the contact wire* includes: 'Abnormal height position', 'Abnormal position in horizontal length', 'Defective movement', 'Incorrect height position', 'Incorrect position in horizontal length', 'Incorrect contact wire position', 'Loose', 'Loose contact wire', 'Droop', 'Incorrect adjustment', 'Height max', 'Height min', 'Horizontal position 200', 'Horizontal position 600', 'Damaged steady tube', 'Steady tube bent', 'Difference between height between hanging posts' and 'Vertical acceleration'. And the category *Damage and wear of the contact wire* was also created. This category includes 'Scorch', 'Defected', 'Defected corrosion protector', 'Defected contact wire', 'Damaged', 'Wear', 'Broken strands', 'Twisted' and 'Spot wear'. More about the different failures can be found in Appendix B.

ArcMap was used in order to get an overview and graphically visualize where different event codes have been logged. By adding the coordinates from the script into ArcMap, the program plotted every coordinate on a table. By adding a map to the table, with SWEREF 99 TM as a reference system, it showed where at the map different event-codes were logged. Data from the all remaining event codes where plotted in different layers together with datapoints from the

infrastructural failure. This was made for all the different infrastructural problems.

3.2. Calculating the relative risk and confidence interval

SQL Server Management Studio was used to sort the data. In all methods, the vehicle data was filtered to only see vehicles in movement, so all vehicles that were traveling at a speed less than 20 km/h were removed. Relative risk, RR, was used to identify an association between the event codes generated by the train computer and infrastructure malfunctions. To calculate RR, a 2x2 matrix was made that included four different subsets of data, se Figure 5. *IE* is the data generated from one event code when there is an infrastructural failure. *IN* is the data from all other event codes except for the investigated event code when there is an infrastructural failure. *CE* and *CN* is the corresponding data when there is not an infrastructural failure. In this case, the used reference data was data generated during two weeks after the infrastructural failure had been fixed.

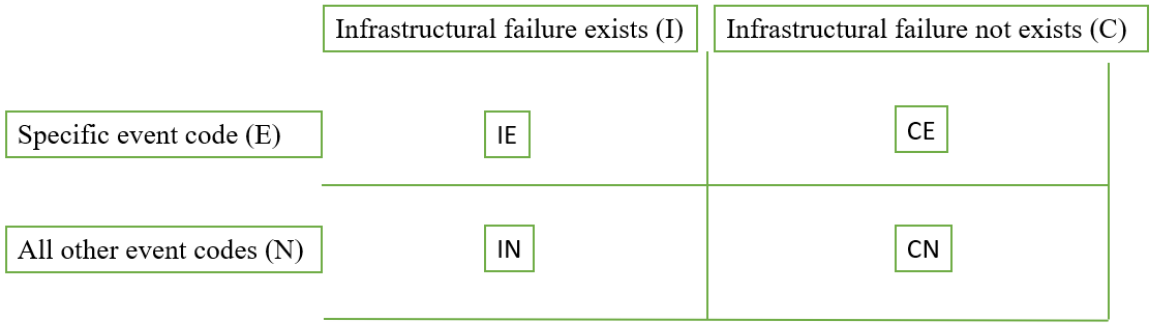

Figure 5 2x2 matrix for calculation of Relative Risk.

To calculate RR the following equation is used:

$$RR = \frac{\frac{IE}{IE + IN}}{\frac{CE}{CE + CN}}$$

In order to create a 2x2 matrix, a table in SQL was created to show event-ID, coordinates, date, and speed for the vehicles, referred as *Vehicle Data* in Figure 6. For Trafikverket's data, a table was created to show coordinates and dates that had an identified infrastructural failure from one of the investigated malfunctions, referred to as *Inspection Data - Failure* in Figure 6. The two tables were joined by date, and distance of 150 m from a track coordinate to a logged vehicle event, referred as *Joined table* in Figure 6. All specific event codes were counted as *IE*, a total was calculated, and *IN* were calculated as total minus *IE*. A new table was then created with the inspection data, but the dates were changed to only see a period of time of two weeks after the infrastructural failure had been fixed. The new table was then joined with the vehicle data and

CE and *CN* were calculated in the same way as *IE* and *IN*. *IE*, *IN*, *CE* and *CN* were then joined by date and RR were calculated for each event code. If an event code occurred in *IE* and not in *CE* it was set that the event code occurred 0,01 times in *CE*. It was set as 0,01 times, to avoid a division by zero.

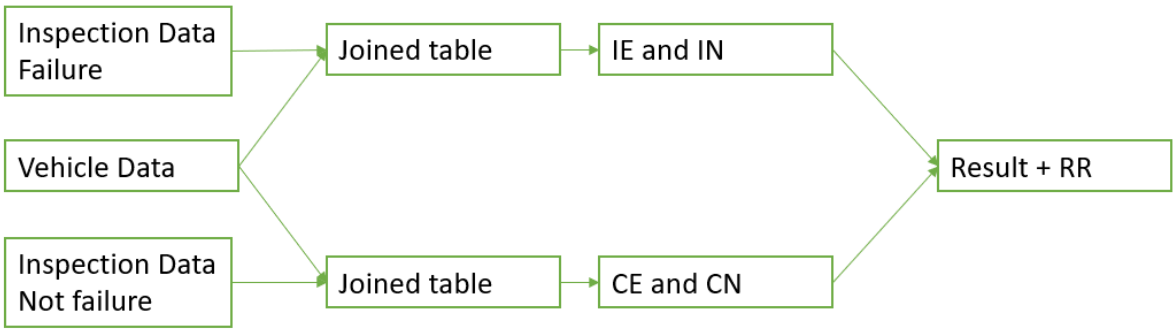

Figure 6 Course of action in SQL.

The null hypothesis is that RR shows a result of one, which means that there is no connection between the infrastructural failure and the event code. The alternative hypothesis is when RR is less than one or if RR is greater than one. If RR is less than one the event codes occur more often when there is not an infrastructural failure, in this thesis these are not further investigated. If RR is greater than one then the risk of triggering the event code is higher when the infrastructural error is present, indicating an association between the two.

In order to decide if the RR-value is trustworthy, a confidence interval was calculated. The interval indicates how far the RR-value can be from the true value, therefore some of the values caused by coincidence can be dismissed. To calculate confidence interval the following equation is used:

$$Confidence\ interval = RR \pm e^{Z_{critical}\sqrt{\frac{1-(\frac{IE}{IE+IN})}{IE}+\frac{1-(\frac{CE}{CE+CN})}{CE}}}$$

This equation is used to show the distribution of the interval. The value $Z_{critical}$ in the equation is calculated by using a table for the degrees of freedom for different percentiles. The degree of freedom is the quantity of independent observations minus the quantity of estimated parameters. Because the degree of freedom in this thesis is very large, the largest value in the table found in Appendix C has been used. The used value is 1000 degrees of freedom. Most statistical tests use a confidence level at 95% which means that there is a 5% chance that the given interval does not contain the true mean value. To lower the risk of having an interval with not true values the chosen confidence interval is higher, 99.9% in this thesis. This gives a better statistical chance to have an interval that includes the true mean values because a 99.9% confidence interval is wider than a 95% confidence interval.

For the event codes that had the highest chance to be able to detect an infrastructural failure, accuracy, precision, and recall. Accuracy is the number of correct predictions divided by the total number of predictions.

$$Accuracy = \frac{IE + CN}{IE + IN + CE + CN}$$

The event codes purpose is not to detect an infrastructural failure. Therefore, the event code will be generated at other points than where there is an infrastructural failure whether it can detect a failure or not. Precision and recall are a way to decide a percentage of how many times the event code was generated when there was an infrastructural failure. Recall is the accuracy for the event code to be generated when there is an infrastructural failure.

$$Precision = \frac{IE}{IE+CE} \text{ and } Recall = \frac{IE}{IE+IN}$$

3.3. Sources of error

The infrastructure data uses a one-dimensional coordinate system to log their data points and the vehicle data uses a two-dimensional coordinate system. In order to join the data in one table the former needed to be altered into a two-dimensional coordinate system. Trafikverket has two-dimensional coordinates for contact wire posts, and this information was used to convert the data points. The contact wire posts are placed at a distance of 60 meters from each other. Therefore, the true position of the infrastructural failure might be 30 meters from the coordinate that has been used.

Data from Trafikverket is collected from measuring vehicles but is also manual input from inspections, therefore different names for the same types of problems may occur. In some cases, there is a lack of information on what the problem and solution to the problem was. This must be considered and is an important factor in the reliability of the data.

All event codes that occurred five times or less were excluded from the list, because it is not likely that these event codes form a pattern if only generated a few times in a three-year span. The remaining event codes were further investigated and all event codes that did not have any connection to the failure that were investigated were removed, some examples of the removed events are ‘incorrect doorstep’, ‘low on fresh water’ and ‘gray water tank almost full’. Some of the removed event codes might have a connection, in this thesis these connections were marked as unlikely.

The purpose of the event codes is to detect vehicle failure or if the vehicle needs maintenance in some other way, like ‘low on supplies’. Some event codes can be generated many times in a short period of time by the same vehicle. This could happen by chance in the investigated area, when we found patterns that hinted that it was clearly one vehicle that had an actual failure with no connection to the track, it was removed. In some case these vehicles may have been missed or data from the vehicle may have had a connection to an infrastructural failure. The basis of the filtering was if only a few vehicles had generated all events in a specific event code, the event code was excluded. If all vehicles, or almost all vehicles, were included but one, or a few vehicles behavior drastically deviated from the others, the deviated vehicle(s) were removed from that event code.

It has not been investigated if all vehicle events were generated by a vehicle during the same day. Therefore, there may be vehicles that do not deviate from the others, but all vehicle events happened in a short period of time. The reason that this has not been investigated is because it is this kind of behavior that might be of interest.

In this thesis, the used reference track was the same track that the failure had occurred at, but after the failure had been fixed. In some case the failure has been set as fixed, but it may not have been the source of the problem that has been fixed which means that the failure could occur again at the same track.

The categorizations of the malfunctions *Incorrect position of the contact wire* and *Damage and wear on the contact wire* has been done with help from documentation from Trafikverket. There may be other data points that could be a part of the created categories, and there may also be other categorizations that are better suited for this purpose.

Like mentioned in *Commuter trains in Skåne and vehicle event codes* some tracks are more used than others and therefore it is reasonable to assume that the event codes are generated more often on these tracks. Therefore, the plotted data may show that the event codes are more frequent on these tracks, but the risk of generated event codes is the same.

3.4. Source criticism

During this bachelor thesis, interviews have been held with Trafikverket, Arriva and Alstom. The interviews have been held with people whose profession correspond to the issue that was discussed during the interview. The company they are working at would not gain from giving us incorrect information and therefore considered credible.

The reports that have been used as references in this thesis are mostly published by Trafikverket. Trafikverket has a role where they educate the public as well as people in the industry. They do not gain from publishing reports without scientific ground. Rail Vehicle dynamics is a textbook made for educational reasons that has been used during this thesis. The book is published by Royal institute of technology, they do not gain for publishing without scientific ground.

4. Result and discussion

This section is based on the results from the methods used from the previous part. All the results have been made by statistical analysis with relative risk and confidence interval.

4.1.1. Cant irregularities

The relative risk in these events vary from 12 to 5 in the presented list. The lowest value in the confidence interval is lower than one for all event codes in the list. This means that there is a risk that there is no connection between the event codes and *Cant irregularity*. After further investigation of the fault messages for all event codes, they were set as unlikely to detect an infrastructure failure. For example, in Table 1 it is shown that the event code "CAN bus once interrupted" have a RR-value of 8. When the event code was further investigated it showed that this event code is generated when, for example, there is failure with the fire detectors, wherefore the event code is not likely to be able to detect a cant irregularity.

Table 1 Remaining faults that are most likely to have a connection with Cant irregularity.

Event code	**RR**	**Confidence interval**		**Fault Message**
		Highest value	**Lowest value**	
2522	12	118	-94	Bogie 1: Axle 1 blocked
17610	8	115	-98	CAN bus once interrupted
1802	7	114	-99	High Voltage Converter module 10 G3 defect
16412	5	112	-101	TCMS: MMI lost communication to the HVA3

4.1.2. Incorrect position of contact wire

The relative risk varies from 1789 to 183 for *Incorrect position of the contact wire*. The event codes that are presented in the list are the event codes that are most likely to be able to detect when the contact wire is in an incorrect position. The lowest value in the confidence interval is higher than one, this means that it is likely that there is a connection between the event codes and *Incorrect position of the contact wire* for all event codes.

Table 2 Remaining faults that are most likely to have a connection to Incorrect position of the contact wire.

Event code	RR	Lowest value of confidence interval	Accuracy	Precision	Recall	Fault message
15014	1789	1683	-	-	-	Current / Voltage Overload Fault
2022	367	261	19%	100%	0.2%	High Voltage Converter A1+A2 off, no DC800V available
15603	252	146	-	-	-	Neutral Section Fault
15507	183	77	-	-	-	Major Traction Control unit fault

As shown in Table 2 the event code *15014* has a very high RR-value, all vehicles had generated this event code, no vehicle deviated in the number of times the event code was generated. This event code is generated when the train is operated in an incorrect manner or when there is damage to the cabling on the train. The event code can also be generated when entering a neutral section without cutting the main switch. The event code is not likely to detect *Incorrect position of the contact wire* based on what the event code can detect. However, more information about where there are neutral sections is needed to assess if the event code is generated because of incorrect operation of the vehicle. As seen on the map to the left in Figure 7 this event code is generated often. There is no clear connection between the location of the generated event code and *Incorrect position of the contact wire.*

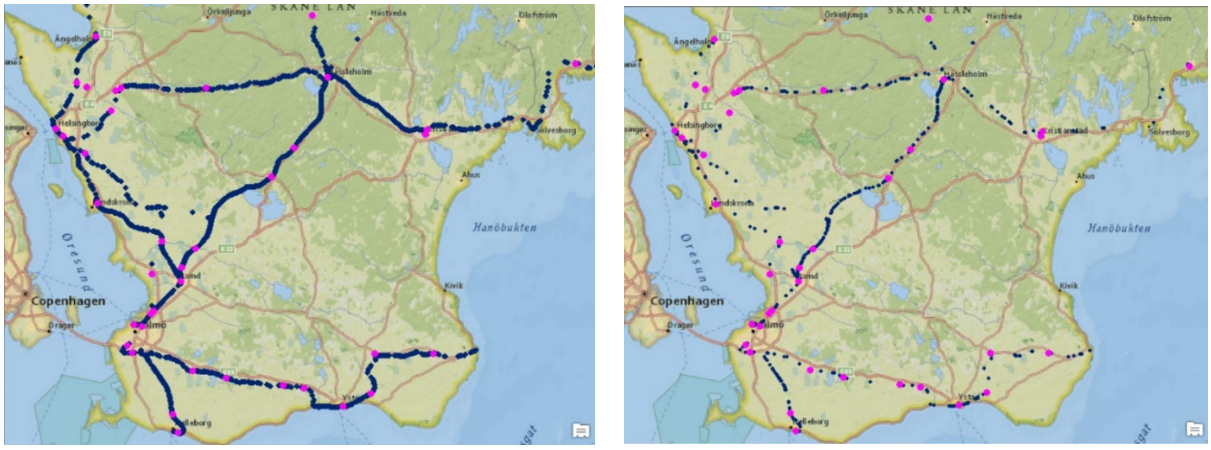

Figure 7 Plotted event codes together with failures connected to Incorrect position of contact wire (event codes in dark blue and infrastructure failure in pink). Left map: Event code 15014, Right map: Event code 2022.

The event code *2022* has a RR-value of 367, all vehicles were represented, and no vehicle deviated in the number of generated event codes. The event code can be generated when for example there is a blown fuse for the auxiliary supply. The part of the train where this event code is monitoring is located where there can be a connection to the contact wire. As seen in Figure 7 this event code is generated a lot on the line Södra stambanan. There is no clear connection between the location of the event codes and the location of the infrastructural failure. The accuracy of the event code is 19%. This means that 19% of the times the event code was generated in a way that can be used to predict an infrastructural failure. The precision of the event code is 100% for the analyzed data and the recall is 0.2%. This means that the event code where generated 0.2% of the times a train passed an infrastructural failure.

The event code *15603* has a RR-value of 252, all vehicles were represented, and no vehicle deviated in the number of generated event codes. The driver should push a button when entering a neutral section to cut off the traction. If the driver pushes the button when there is no neutral section, the event code is generated. The left map in Figure 8 shows the event code plotted together with all faults in the category *Incorrect position of the contact wire.* The event code seems to have a connection in the plotted data, however this event code is manually started and can therefore not be an indication of that an infrastructural failure has occurred. This event code needs to be further investigated with information on where there are neutral sections to see if the vehicle is operated in an incorrect manner or if this event code is generated where it should not have. If the event code is generated where there is a neutral section, this may depend on an infrastructural failure.

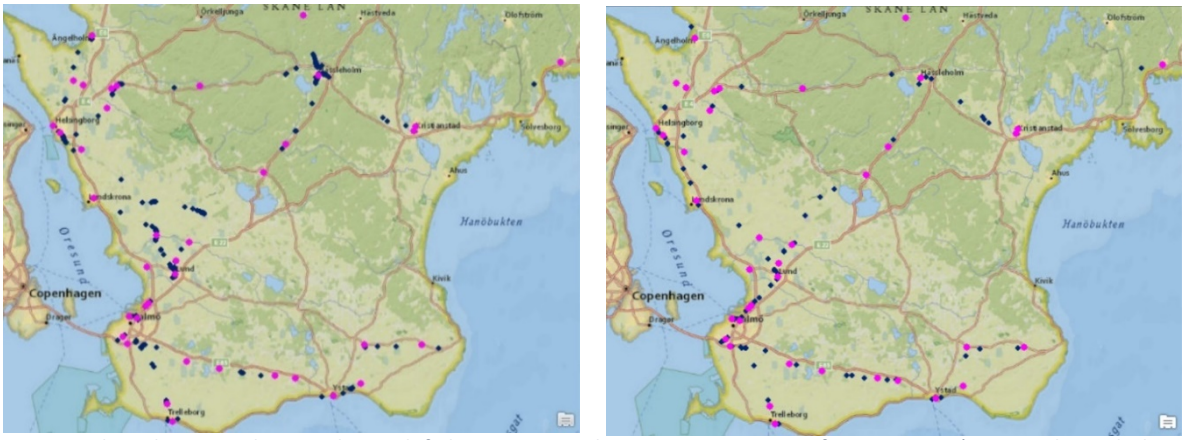

Figure 8 Plotted event codes together with failures connected to Incorrect position of contact wire (event codes in dark blue and infrastructure failure in pink). Left map: Event code 15603, Right map: Event code 15507.

The last event code in the list is event code *15507*, it has a RR-value of 183. All vehicles were represented, and no vehicle deviated in the number of generated event codes. This event code is generated when the traction unit group is switched off. Like event code *15603,* this event code is manually started, therefore this event code is not an indication that there is a fault connected to the contact wire, however there can be a connection to why the driver chooses to switch off one off the traction unit groups. The map to the left in Figure 8 shows the event code *15507* together with all faults in the category *Incorrect position of the contact wire.* The location of the event code and the location of the infrastructural failure seem to have a connection, but like mentioned, these event codes are manually started and therefore the true location where something happened that made the driver take these measures may be at another location.

4.1.3. Damage and wear on contact wire

As shown in Table 2 and Table 3, two of the event codes presented are the same, *2022* and *15603*. The failures in the category is related to each other since they both occur on the contact wire. The event codes presented in the list are the ones that are most likely to have a connection to *Damage and wear on the contact wire*. The relative risk varies from 324 to 60 for the event codes connected to *Damage and wear on the contact wire*.

Table 3 Remaining event codes that are most likely to have a connection to Damage and wear on the contact wire.

Event code	RR	Lowest value of confidence interval	Accuracy	Precision	Recall	Fault message
2005	324	218	5%	100%	0.3%	DC110 V supply: Earth fault level 1
15603	220	114	-	-	-	Neutral Section Fault

2022	108	3	4%	100%	0.1%	High Voltage Converter A1+A2 off, no DC800V available
2006	60	-46	-	-	-	DC110 V supply: Earth fault level 2

Event code *2005*, "DC110V supply: Earth fault level 1" is the code with the highest RR-value. This failure has two event codes connecting to it depending on how big of a leakage there is, both presented in Table 3, though the event code *2006* has a confidence value that indicates that this event code may not have a connection to the infrastructural failure. The event code *2006* indicates that there is a failure after 15 minutes, which means that the location where this event code is logged is not where the event code first started, therefore the found connection is coincidental. Both event code *2005* and event code *2006* occur mainly on Västkustbanan and Trelleborgsbanan, see Figure 9. The event code *2005* occurs a lot, both when there is an infrastructural failure and when there is no infrastructural failure. There is no clear connection between the location of the infrastructural failure and the location were the event code was generated. Both event codes occur were there are a lot of infrastructural failures, like in the area between Malmö and Lund. This area is heavily trafficked so there is reason to think that some event codes triggers more often in this area because of that. Event code 2005 had an accuracy of 5%, a precision of 100% and a recall of 0.3%.

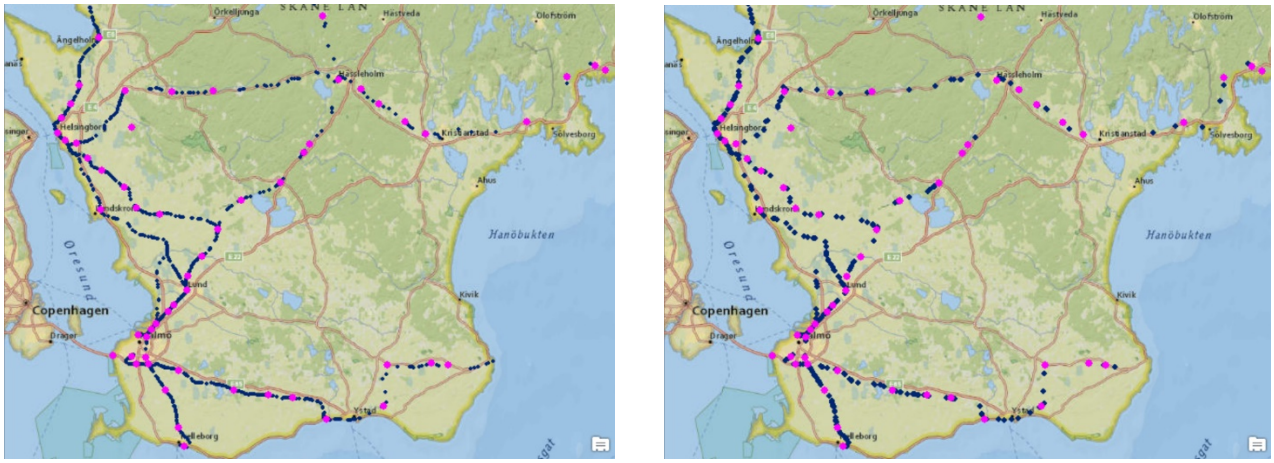

Figure 9 Plotted event codes together with failures connected to Damage and wear on the contact wire (event codes in dark blue and infrastructure failure in pink). Left map: Event code 2005, Right map: Event code 2006.

The event code *15603* is generated by incorrect operation of the vehicle. This event code is connected to where the neutral sections are located and should be further investigated with information on their location. There may be a connection between the event code and the infrastructural failure if there is a neutral section in the location. On the map to the left in Figure 10 it can be shown that the location of the infrastructural failure is close to the location of the generated event code.

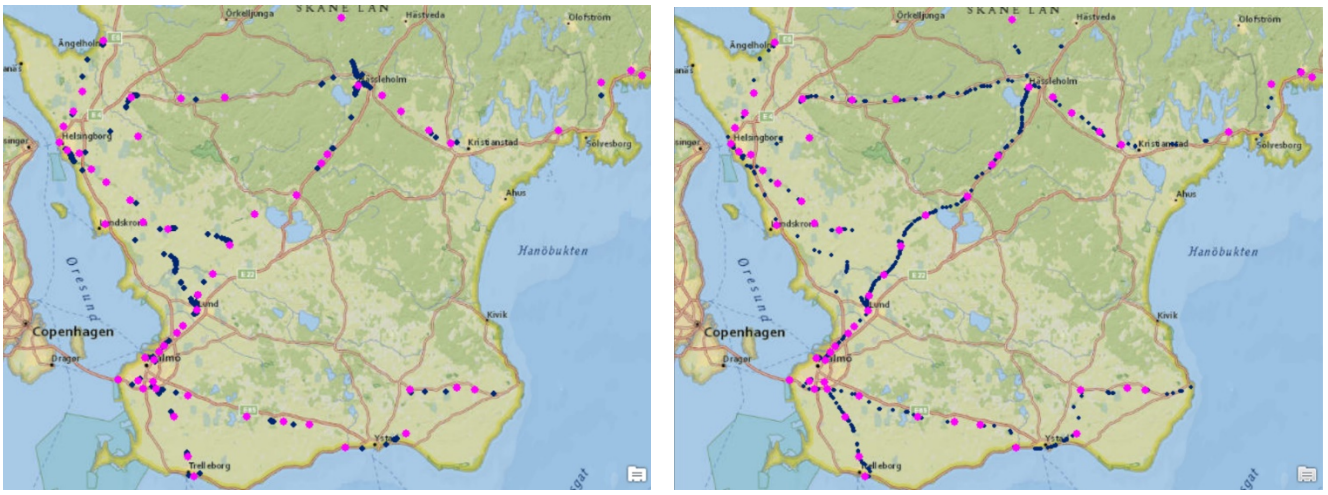

Figure 10 Plotted event codes together with failures connected to Damage and wear on contact wire (event codes in dark blue and infrastructure failure in pink). Left map: Event code 15603, Right map: Event code 2022.

The event code *2022* is mentioned in the result for *Incorrect position of the contact wire* as well. This event code can be generated by a blown fuse for the auxiliary supply. This event code occurs both when there is an infrastructural failure and where there is no infrastructural failure. There is no clear pattern between the location of an infrastructural failure and the location of the generated event code. The event code had an accuracy of 4%, a precision of 100% and a recall of 0.1%.

5. Conclusion and future development idea

In conclusion there is no connection between the event code presented for *Cant irregularities* but there may be a connection between some of the event codes presented for *Incorrect position of the contact wire* and *Damage and wear on the contact wire*. For *Incorrect position of the contact wire* and *Damage and wear on the contact wire* the event code that is most likely to have a connection is *2022* because of what the event code is indicating, however there is no clear connection between the location of the failure and the event code in the plotted data. Accuracy for the event code is between 4% and 19%. This means that 4-19% of the times the event code was generated in a way that can be used to predict an infrastructural failure. The precision is 100% in the investigated data because the event code did not occur in the reference data. Recall value for event code 2022 is between 0,1% and 0,2%. It means that it is 0,1-0,2% chance that the event code is generated when there is an infrastructural failure. Even though the precision is 100% the accuracy is between 4-19% and the recall is only 0,1-0,2%. This means that the event code would not generate most of the times the trains passes an infrastructure failure. The event code *15507* is another event code that may have a connection to *Incorrect position of the contact wire*, this event code is manually started and therefore a solution to a problem that occurred on the vehicle. The problem that occurred may have a connection to the infrastructural failure. Because this code is manually started the event code cannot be used to predict an infrastructural failure but may still have a connection to why the event code is occurring. The event code *15603* may have a connection to infrastructural failures in the categories *Incorrect position of the contact wire* and *Damage and wear on the contact wire* but this event code is also manually started and can therefore not by itself be used when predicting infrastructure failures. It may be possible to use the event code *15603* together with information about neutral sections to predict an infrastructural failure.

For *Damage and wear on the contact wire* the event code *2005* may have a connection to the infrastructural failure based on the RR-value and the reason why the event code is generated, but the event code has no clear connection in the plotted data. Accuracy for event code 2005 is only 5%. The precision is 100% in the investigated data because the event code did not occur in the reference data. Recall value for event code 2005 is 0,3%. It means that it is 0,3% chance that the event code is generated when there is an infrastructural failure. This means that this event code is not likely to detect an infrastructure failure.

All listed event codes need to be further investigated to see if the connections that has been found are possible. However, if a failure could be detected on the tracks by the vehicles, it is suggested that an automatic alert should be sent to the train dispatcher from Alstom. The algorithm that triggers the automatic alert should be structured in a way that the sensors need to be generated a certain

amount of times by a certain number of vehicles within a specific distance to reduce false alarms. The dispatcher can then take necessary precaution depending on the severity of the failure.

5.1. Future development ideas

In this thesis, all event codes have been considered individually and not in combination with each other. There may be a pattern in a combination of event codes if the data is analyzed per train and not per event code. To analyze all event codes, it is no longer possible to use the statistical methods that have been used in this thesis because if the data is sorted by train and not event, each data point needs to have information about all event codes. Therefore, each data point will have more than 735 variables since there are 734 different event codes that have been generated during the investigated time. Machine learning or clustering methods need to be used to analyze data with that many variables.

If all vehicles had been investigated initially to remove vehicles with known faults, the outcome may have been different. In this case all investigated vehicles could be seen as vehicles without malfunctions. To further make the data more reliable a reference track without any malfunctions could be used.

Like mentioned in the Conclusion and future development idea information about the track could have led to a more reliable result when analyzing the data, specifically information on where neutral sections are located.

As mentioned in Introduction and background, it is possible for vehicles to detect the same type of problems that the measuring vehicle can detect if they are carrying the right equipment.

Bibliography

Alstom. (2016). *Transport services for your rail system.* Alstom.

Andersson, E., Berg, M., & Stichel, S. (2014). *Rail Vehicle Dynamics.* Stockholm: Kungliga Tekniska Högskolan.

Arriva. (01 03 2020). *Om Arriva*.Arriva

Banverket. (2006). *Lärobok kontaktledning.* Banverket.

Corshammar, P. (2006). Södra stambanan Malmö - Stockholm Sveriges viktigaste järnväg. Stambanan.com

De Rosa, A., Kulkarni, R., Qazizadeh, A., Berg, M., Di Gialleonardo, E., Facchinetti, A., & Bruni, S. (2020). *Monitoring of lateral and cross level track geometry irregularities through onboard vehicle dynamics measurements using machine learning classification algorithms.* Department of Mechanical Engineering, Politecnico di Milano and Department of Aeronautical and Vehicle Engineering, KTH Royal institute of Technology.

Finn, V. (2019). *TTT – Tillsammans för Tåg I Tid. TTT*

Gummeson, M. (2019). Tillsammans för tåg i tid Resultatrapport 2019. TTT

Hjort, J. (2010). *Kontaktledningsfel upptäckta vid mätning av kontaktledning.* Trafikverket.

Honauer, U., & Ödeen, S. (2019). Underhållsplan 2019–2022. Trafikverket

Karis T; Berg, M; Stichel, S; Li, M; Thomas, D; Dirks, B. (24 11 2017). Correlation of track irregularities and vehicle responses based on measured data. Stockholm Kungliga Högskola

Lundmark, J. (2007). Rail Grinding and its impact on the wear. Luleå University of Technology

Olsson, A. (05 03 2020). (L. Rydstedt, & C. Åkerlund, Intervjuare) Arriva

Skånetrafiken. (10 03 2020). *Tidtabeller*. Skånetrafikens

Trafikverket. (28 06 2019). *Sveriges Järnvägar*. Trafikverket

Trafikverket. (Utgåva 2.0). *OPTRAM kontaktledning - Dynamiska anmärkningar script .* Trafikverket .

Trafikverket, Region Stockholm, MTR. (2019). *Rapport Senior Samverkan: Från ord till handling - åtgärdet för punktligare pendeltrafik.* Stockholm Stad.

t-Table. (2007).

Weston, P., Ling, C., Goodman, C., Roberts, C., Li, P., & Goodall, R. (2007). *Monitoring lateral track irregularity from in-service railway vehicles. University of Birmingham, Loughborough University*

Weston, P., Roberts, C., Yeo, G., & Stewart, E. (2015). *Perspectives on railway track geometry condition monitoring from in-service railway vehicles. University of Birmingham.*

Appendix A

Table 4 List over event code that has been figured during this thesis.

Event code	*Fault message*
2522	Bogie 1: Axle 1 blocked
	This event code usually triggers when there is a failure with the breaks. When the event code has been generated maintenance operators usually check for flat spots on the wheels or if the breaks are locked.
16412	TCMS: MMI lost communication to the HVA3
	The climate control unit does not have a power supply or the communication with the device is lost. The event code usually triggers when there is a software update or when the automatic fuse is switched off.
1802	High Voltage Converter module 10 G3 defect
	High voltage converter module creates 800V DC to the three-phase inverter and to the battery charger. This module is connected to the lightening protector and the event code can trigger when there is overvoltage or when there is a rush in the traction.
17610	CAN bus once interrupted
	This event code contains a collection of different malfunctions. It can be generated when there is a problem with the fire detector or that the detector is damaged. It can also detect when there is a problem with the connection to the train computer.
15014	Current / Voltage Overload Fault
	The event code can trigger when the train is operated in an incorrect manner. It can also trigger when there is a damage on the cabling.
15507	Major Traction Control unit fault
	"Self-test" not approved or that the unit is off. The traction unit group is switched off, therefore reduced traction. This code is manually started. This event code is usually occurring in a combination of other event codes.
2005	DC110 V supply: Earth fault level 1
	The primary power supply for the low voltage devices, like breakage and traction devices, are 110V DC. The event code triggers when the earth fault monitoring shows that there is an earth fault. This event code occurs in two different levels where level 1 is the least critical.
2006	DC110 V supply: Earth fault level 2
	Se event code *2005.* The event code levels depend on how big the leakage is, where level 2 is the most critical. This event code is generated after there has been an earth fault for 15 minutes.
15603	Neutral Section Fault
	The driver manually pushes a button before entering a neutral section in order to avoid damage to vehicle and infrastructure. This event code is generated with incorrect use of the button.
2022	High Voltage Converter A1+A2 off, no DC800V available
	The high voltage unit that transforms 1800V DC from the traction to 800V DC to the batterie inverter 110V DC supply and three phase supply. Converter switched off, no supply from the traction or that the traction is broken. This code can be generated when there is a blown fuse for the auxiliary supply.

Appendix B

Datapoints – Incorrect position of contact wire:

Trafikverkets Swedish name	**English translation**	**Description**
Avviknade höjdläge	Abnormal height position	Weight too low or too high (in these cases too low)
Avvikande sidoutslag	Abnormal position in lateral length	Position of contact wire incorrect.
Felaktigt höjdläge	Incorrect height position	Weight too low or too high
Felaktigt sidoläge	Incorrect position in lateral length	Contact wire in wrong lateral position
Felaktigt trådläge	Incorrect position of contact wire	Contact wire in wrong position
Höjd max	Height max	Maximum height for contact wire
Höjd min	Height min	Minimum height for contact wire
Sidoläge 200	Lateral position 200	Contact wire is too close to the center of the track.
Sidoläge 600	Lateral position 600	Contact wire is too far from the canter of the track.
Utspetsning	Difference between height to high or too low.	Contact wire too high or too low in a specific range.

Datapoints – Damage and wear on contact wire

Trafikverkets Swedish name	**English translation**	**Description**
Brännskadad	Scorched	Contact power line scorched
Defekt	Defected	Defected contact wire
Defekt korrosionsskydd	Defected corrosion protector	Defected corrosion protector
Defekt lina	Defected contact wire	Defected corrosion protector
Skadad	Damaged	Contact wire damaged
Slitage	Wear	Wear on contact wire
Punktslitage	Spot Wear	
Krokig/Vriden	Twisted	Twisted Contact wire
Kardeler av	Broken strands	Strands on the contact wire is broken
Vertikal acceleration	Vertical acceleration	Vertical acceleration in contact wire

All sorting has been made with help of documents collected from Trafikverket. These documents are Textbook about the overhead line (Banverket, 2006), Dynamical remarks (Trafikverket, Utgåva 2.0) and Faults appearing on the overhead line detected when measuring the overhead line (Hjort, 2010).

Appendix C

Table 5 Values for various confidence levels with various degrees of freedom (t-Table, 2007).

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										