Learning Patterns in Train Position Data

Classifying locations by identifying station specific patterns

Ivan Smilenov[1]

Supervisor(s): Prof.Dr. M.M. de Weerdt[1]**, I.K. (Issa) Hanou**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Ivan Smilenov
Final project course: CSE3000 Research Project
Thesis committee: Prof.Dr. M.M. de Weerdt, I.K. (Issa) Hanou, Asst. Prof. Jing Sun

Abstract

Solutions for the Train Unit Shunting Problem are constantly being researched and improved to become more efficient and match the needs of train transport in the Netherlands. For this reason, we are exploring new ways to find patterns in the train data to identify where those solutions could be enhanced. More specifically, we are trying to find patterns that make identifying different locations possible. We identify patterns in the capacity of the shunting yards and the types of trains used in various locations, which result in reasonable accuracy in classification. Some locations operate closer to their capacity, and some require longer paths to get inside the shunting yards. These findings could be helpful not only for the planning algorithms but also in identifying which locations might need expansion or restructuring and where more of the train fleet should be allocated.

1 Introduction

In densely populated countries like the Netherlands, train transport is prominent and has a lot of traffic. This requires very complex planning of the allocation of trains and resources since the infrastructure is very limited compared to the significant demand. Additionally, depending on the different demand throughout the day, a varying number of trains are needed. Because of this, some of them should stay parked somewhere for most of the day without disrupting the traffic. This is where the Train Unit Shunting Problem appears. The aim is to create a feasible schedule given different constraints like track availability, needs for servicing, etc.

Much research has been done in the field to find a good solution for the problem [1], [3]. Currently, there are two main types of algorithms used by Nederlandse Spoorwegen (NS) - one using Deep Reinforcement Learning (DRL) and one using local search [1]. Most of them still rely on human input to provide feasible solutions. The algorithms that do not rely on that have other disadvantages, like infeasible computation time or incompleteness- missing activities in the schedule. Furthermore, the solutions created by such algorithms may be inconsistent, which is generally not preferred in scenarios like timetable scheduling [5]. Because of these reasons, we are still in search of an optimal algorithm that is adaptive to different scenarios and computes in a reasonable time. This is why the local search using different heuristics is in focus. Here lies the reason why studying and finding patterns in already existing solutions, i.e. schedules, could be very beneficial to improving those algorithms in the future and limiting the need for human input. Using the insight we get from these patterns, different aspects of the existing algorithms can be improved, such as the heuristics used in local search.

To the best of our knowledge, there has not been in-depth research into finding patterns that arise in different locations. Our research aims to bridge this gap by finding such patterns in existing solutions and using them to differentiate between locations automatically as a verification method.

With the help of this research, we can identify different patterns in train shunting, given the scheduling of the train movements in a specific station. By finding these patterns, we can automatically detect when a solution is coming from one station or another. Identifying which patterns arise in which locations can help optimise future and existing algorithms when we fine-tune the specific heuristics used in them or identify best practices. Additionally, these patterns can help when designing the locations by identifying which need expanding or which layouts are better performing and easier to create schedules for (shuffleboard[1] or carousel[2]).

The rest of this paper is structured as follows: Section 2 gives background on the main methods, data and problems and briefly discusses related work. In section 3, we describe the research process step by step. Section 4 describes the achieved results from the conducted experiments. After that, there is a section where we reflect on the ethical side of the research and whether it is reasonably reproducible. There is a separate section where a discussion is done on the meaning of the achieved results. The last section contains the conclusion on the research and ideas for possible future work.

2 Background

2.1 Train Unit Shunting Problem

The Train Unit Shunting Problem (TUSP) is a complex logistical challenge railway operators face involving the efficient organization and management of train units (rolling stock) in and around railway yards. First introduced by Freling et al. [4], TUSP is still the main topic of broad research, and the algorithms used for solving it are constantly being improved to match the increasing demand. The main problem consists of multiple sub-problems like the matching, track assignment, shunting routing and shunting maintenance problems [7]. All of those are NP-Hard problems, which are the focus of different optimizations. As mentioned in the previous section, NS used to develop a scheduling algorithm using deep reinforcement learning first introduced by Peer et al. [10]. This algorithm was shown to provide outstanding solutions, which required much less training than others. Furthermore, it maintained reasonable consistency, which is a significant factor in scheduling. However, Gevel [5] argues that this algorithm sometimes provides partial solutions, which are impossible to use unless completed, and that is generally not a trivial task. For this reason, the proposed local search algorithm [1] is still preferred by NS. The solution is represented as a graph where the nodes represent the different types of activities performed. The algorithm starts with an initial solution and uses simulated annealing[3] to iterate over better solutions. This results in a significantly slower computation time than the DRL algorithm, but it is currently still feasible. Furthermore, it is easier for planners to work with, which is a crucial aspect of these algorithms [6]. The long computation time begs for the optimization and employment of different heuristics.

[1] Tracks with a dead-end

[2] Tracks are accessible from both ways

[3] Technique for approximating the global optimum of a function

2.2 Clustering and Classification

Clustering algorithms play a crucial role in computer science, data science, machine learning, pattern recognition, and many other fields. While numerous clustering techniques and algorithms exist, our research focuses on two specific methods: K-means clustering and the Random Forest Classifier. We chose these because we have a precise number of clusters, which is usually not the case. Furthermore, they are speed-efficient and not very sensitive to outliers.

K-means is one of the most well-known and widely used clustering algorithms. The primary goal of k-means clustering is to divide a set of n data points into K clusters, where each cluster is defined by its centroid. The algorithm follows a straightforward process of iteratively selecting initial centroids, assigning data points to the nearest centroids to form clusters, recalculating the centroids by calculating means, and repeating these steps until the centroids stabilize or a set number of iterations is completed. If those clusters are well-defined, the results will be very effective. K-means works particularly well with numerical data, which often requires normalization. Various normalization techniques are available, and selecting the most suitable one depends on the specific dataset [2], so it is best to experiment with multiple. New data points are classified by evaluating their distances to the cluster centres. Visual verification of the clustering can be straightforward by performing Principal Component Analysis (PCA) on the data and labelling the points according to their assigned clusters. PCA is a standard tool used to reduce the dimensionality of complex data, thus simplifying its structure and finding underlying patterns [12].

The Random Forest Classifier (RFC) is a versatile machine-learning algorithm known for its robustness and accuracy. It is mainly used for classification tasks but can also be used for clustering. The key advantage of Random Forest is that it consists of an ensemble of decision trees [8], each trained on a subset of the data. The final classification is determined by aggregating the predictions from all individual trees. That improves overall accuracy in predictions and controls overfitting. In a clustering context, RFC can be used to measure the similarity between data points based on the frequency with which they end up in the same leaf node across all trees in the forest. This similarity measure can then be used to group data points into clusters. Random Forests have several advantages, like handling large datasets with higher dimensionality and providing estimates of feature importance. This can be valuable for understanding the underlying structure of the data, such as existing patterns.

2.3 Related Work

As mentioned earlier, there is no extensive research in the aspect we are considering and with similar data. However, we can draw inspiration from other attempts to perform classifications to draw conclusions about underlying patterns. Automatic classification has been employed in various fields in recent years. However, there are many aspects of that process that little attention is paid to or disregarded altogether. In Luz et al. [9], classification is utilized to detect heart diseases from electrocardiograms (ECGs) as this is the most widely used non-invasive method for testing. This work mentions a crucial aspect of the classification that also affects our research- the data preprocessing. One of the main points is that no matter how the data was collected, there is always the possibility of noise or inaccuracies, which should be eliminated in the best possible way. We will later see that it is especially true regarding GPS coordinates and approximations like track assignment in our data. Another critical aspect they mention is ensuring the training and testing data is as unbiased and inclusive as possible. Data from all classes should be provided with equal weights to ensure realistic accuracy. We want to expand this statement even more; picking the features can also negatively influence absolute accuracy if we identify 'fake' trends or ones that do not contribute to our findings in a useful way. For example, GPS coordinates can be used to classify locations with 100% accuracy, but no insights are gained from that. That is why we put a lot of effort into picking the right features.

2.4 Train Position Data

For the goals of this research, ProRail provided us with a dataset containing train position data. The trains in question are commercial trains operated by NS, the primary Dutch railway operator. This data was collected from May 2023 to February 2024 in seven stations across the Netherlands. The data contains the trains' GPS locations, timestamps, and labels like the type of activity[4], track assignment, and many more features that are not strictly relevant to our goals. Additionally, multiple versions of the data were provided with different amounts of processing since the GPS locations and track assignments are not always accurate. Code was also available to fetch the data from Azure Blob Storage and visualize it for a specific location and timeframe.

3 Methodology

Having access to the relevant data, together with our research group, we designed a data structure that contained only the data we deemed relevant at that point. It was structured in a way that we combined the whole paths of the trains instead of having only different timestamps with locations. The information about the paths contains the type of activity, the type of the train, and the list of visited tracks with the corresponding timestamps. This data structure enables further processing and extracting more valuable features when the solutions are considered as a whole. The other essential part of this data structure is the addition of a filtering of the trains to only those that enter the shunting yards at some point. That is done by ensuring the specific train has been on tracks inside the shunting yard for at least three[5] consecutive timestamps. The reason for the filtering is that we focus mainly on investigating patterns of the movements inside the shunting yard. We are aware that movements inside and outside are not entirely independent, but from now on, we disregard that fact.

Using the completed data structure, we examine the data and determine which specific features could be useful to in-

[4] long stop, short stop, shunting, entering, exiting

[5] That is done to eliminate the possibility of inaccurate track assignment in the data

vestigate and which others we could extract from the available data. First, we focus on separate train paths inside the yards instead of the whole solution[6]. This can show whether different single units follow different patterns in different locations. In this case, we can group the features into three categories:

1. Geographic features
2. Time-specific features
3. Path-specific features

The first includes features such as compass direction, GPS locations, etc. These can be very easy to classify, but we will avoid using them because they do not provide any insight into our ultimate goal. The time-specific features like relative time of entry in the yard[7], time spent inside the shunting yard did not provide any useful patterns, keeping in mind that we only look at trains separately at this point. The third kind of features include how many tracks the train covered to get to the yard, inside the yard, and so on. These features already provided some visible patterns and could be used for classification, combined with some time-specific features.

Since we want to classify a specific number of different locations and with the possibility of adding new ones, we will investigate the two well-known algorithms mentioned previously- Unsupervised K-means [13] and RFC. K-means is featured because of its ease of use and its combination with PCA, which provides visual assurance of whether the clustering is what we expect. RFC, on the other hand, is very versatile when it comes to multiple features, and it can be used to identify which of these features are useful and contribute to the classification. The latter proved to give better results and insights complimenting our assumptions, which will be stated in the next section and discussed later.

After that, we focus on the solutions for whole days rather than for separate trains. Here, we investigate features like what percentage of the trains are of a specific type, what percentage of the location's capacity is used, how many trains have been stopped for a long time, the amount of manoeuvres performed, etc. We chose these features because we suspect they would give the most clear indication of whether there are patterns in the usage and layouts of the stations. Again, we use the earlier algorithms to evaluate whether the patterns we identify statistically in those features can be used for classification.

4 Experimental Setup and Results

For the experiments performed in this section, data from six locations is used from May 2023 till the end of February 2024. The data from Arnhem Goederen was not used because the traffic there was very limited, and filtering the tracks belonging to the shunting yard was unreliable. We investigate only trains that, at some point, stop at the shunting yard. Transiting trains are ignored. In table 1, we show the capacity of each location in number of trains and the number of trains that have passed in that period. It is important to note that capacity is an imprecise term- it depends on multiple factors, such as the length and type of the trains present or varying parking strategies. That is why we assume the total capacity is the most trains that have been present in the shunting yard over the whole period. An indication of that imprecision is Dordrecht, where the capacity is much smaller compared to the others. When observing the traffic visualisation, the reason is apparent- trains are often parked inside the station instead of the shunting yard. However, we continue working with the same assumption as the shunting yards are the main focus of this research.

Location	Capacity	Number of Trains
Arnhem	19	5922
Watergraafsmeer	32	6355
Amersfoort	46	10306
Dordrecht	5	1783
Utrecht	40	11233
Hoofddorp	32	42493

Table 1: Train station capacity and traffic

4.1 Classifying single train paths

As mentioned in the previous section, we first investigate whether the single train paths follow any patterns. We start by extracting new features from the data in the three categories we identified earlier. After that, we plot different features and try to identify any visual patterns. For example in fig.1 we plot the average number of tracks visited throughout the path of the train[8] in the whole station. We can already suspect that in some locations, trains almost immediately go inside the yard and out of it without any additional manoeuvres, which is not the case in others. We plot the average entry lengths in fig.2 to verify whether that is the case. Only four of the six stations are used for this experiment because Arnhem and Watergraafsmeer have roughly the same traffic, and Arnhem has a little less capacity. Amersfoort and Utrecht have roughly the same traffic and capacity. That means we can compare them pairwise. We immediately notice that even though stations are of similar capacity and traffic, trains there have very different lengths of entry into the shunting yard.

Having established that this could be helpful for the classification, we run K-means to see whether our assumptions are correct. We run that with four features- the whole path length, the path length inside the shunting yards, the path length before entering the yards, and the number of U-turns[9]. We chose this last feature because we suspect the different layouts might influence the number of manoeuvres. Even though the clustering seems well separated in fig.3, the accuracy[10] we

[6] The combination of all the train movements inside

[7] That is the time passed since the train entered the station's whereabouts

[8] We randomly sample an equal amount of such different paths from each dataset

[9] We assume a train does a U-turn if its direction changes by more than 90 degrees. Alternatively, it starts going in the opposite direction

[10] We assign the clusters to the class with the most occurrences inside

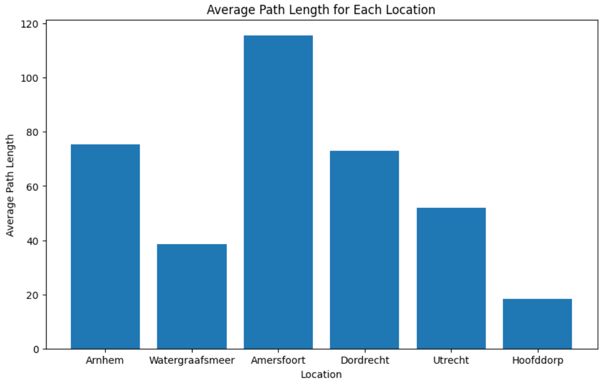

Figure 1: A bar plot showing the average path number of tracks visited by trains across the locations

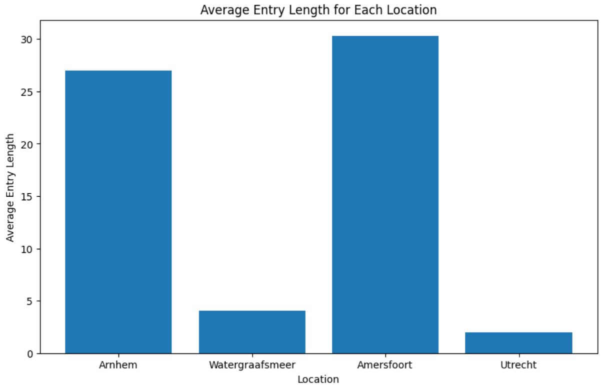

Figure 2: A bar plot showing the average number of tracks visited before entry in the shunting yard

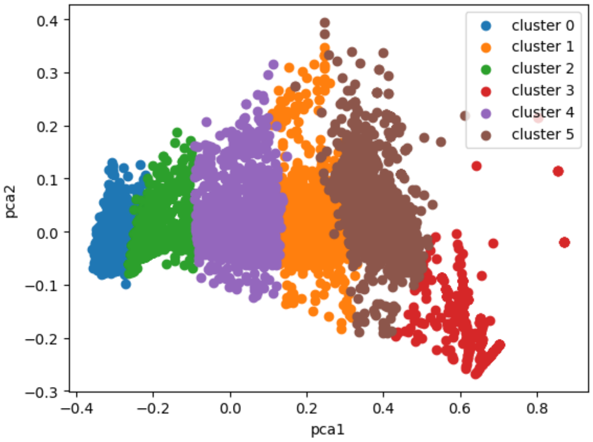

Figure 3: A scatterplot showing the clustering of the data points after performing PCA

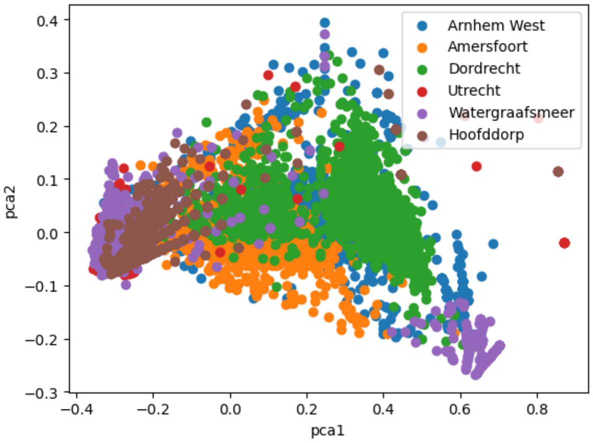

Figure 4: A scatterplot showing the true labels of the data points after performing PCA

achieve is only 66%. When we examine the actual labels of the points in fig.4, it is clear why we have lower accuracy. It is apparent that these features are not ideal for classification with K-means.

Using K-means poses some difficulties because it requires normalising the features to work optimally [2], which can be challenging if we want to use features of entirely different magnitudes, like timestamps. This is why we look into RFC when we try to incorporate those. Aside from the mentioned earlier four features, we add three timestamps - the time of first entry in the shunting yard, the time when the train leaves the shunting yard, and the duration of time spent inside the yard. The times here are relative to the first timestamp the specific train enters the vicinity of the station. That is done to ensure that these times are comparable across trains. Another advantage of the RFC we are going to employ is the feature importance estimation [14]. This means we can see which features contribute to the classification and focus more on them in the search for patterns, as seen in fig.5. In this case, we can see that inside[11], entrylen[12] and Pathlen[13] are helpful for the classification, which proves the earlier observations about the lengths of the paths. In contrast, the others do not seem to contain any underlying patterns. For example, if we examine the number of U-turns in fig.6, there is no clear distinction between the different locations. That shows that no matter what the layout is (shuffleboard or carousel), the trains perform roughly the same amount of manoeuvres; hence, not many trains block others. From these findings, we can also deduce that we should not focus on the times of entry/exit in the shunting yard since they do not contribute significantly.

Using this method, we already achieved about 86% accuracy, which is a clear improvement over the first one. The evaluation is done using 70% of the data for training and the rest for testing. Additionally, we use Randomized Search[14] to find the best parameters for the classifier [11].

[11] The length of the path inside the shunting yard only

[12] The length of the path before entering the shunting yard

[13] The total length of the path inside the station

[14] This method explores random combinations of hyperparameters in a specified range or distribution

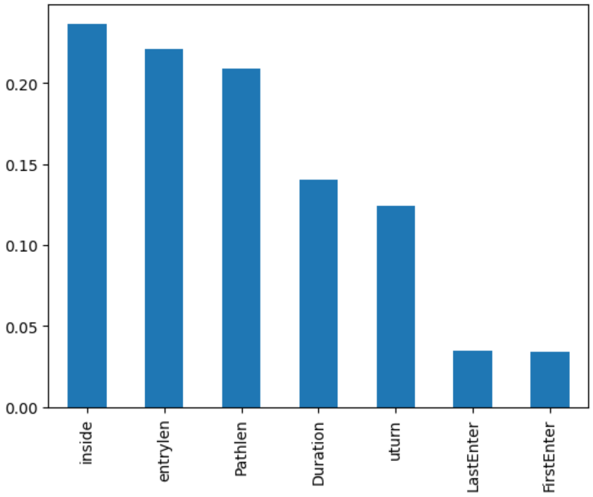

Figure 5: A bar plot showing the importance of each feature. The path lengths have the highest significance

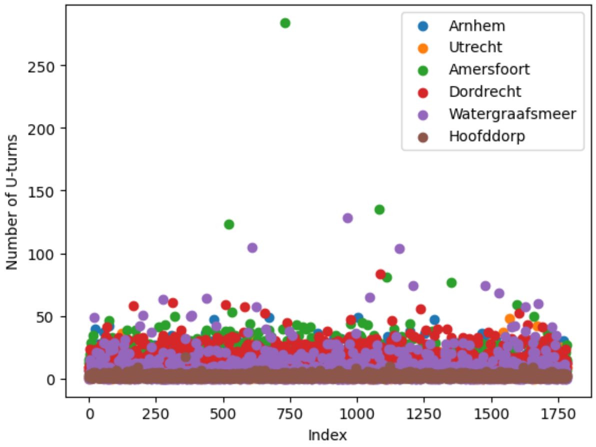

Figure 6: A scatterplot showing for each station the number of U-turns the trains perform inside

4.2 Classifying whole solutions

Even though we achieved decent accuracy in the classification, we still have not identified very clear patterns in the schedules that could be reasoned about apart from the path lengths. That is why we are shifting our focus to whole solutions and not on single entities anymore.

First of all we state our definition of a solution: All train movements inside the region we investigate for a period of 24 hours. According to experts, planners use the period from 8 a.m. to 8 a.m. the following day, which will also be our assumption. Using this information, we now use the same data but aggregate it across days and end up with entirely new features from what we looked into in the previous subsection. We focus primarily on the capacity and train distribution over time. Considering this, we arrive at the following set of features:

- Percentage of the capacity of the yard used
- Number of trains parked for more than 16 hrs[15]
- The ratio of trains parked over a long period of time and the total number of parked trains
- What percentage of the trains each type accounts for

By plotting these features, we can identify that some shunting yards operate closer to their capacity than others. We can see evidence of that in fig.7. Most of the time, Utrecht operates closer to its total capacity than Amersfoort. The latter usually operates at 0- 40% of its total capacity and Utrecht around 20- 70%.

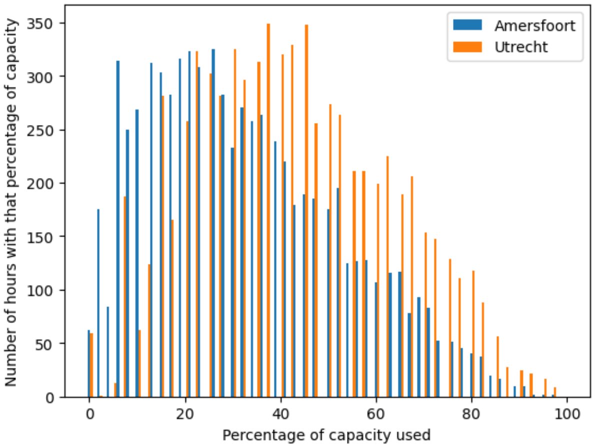

Figure 7: A bar plot comparing how much of the time a certain percentage of the capacity is used in Amersfoort and Utrecht

Even though we are aware that NS has different amounts of trains of different types, those do not always follow the same distribution across all locations. As an example, the largest amount of trains are SNG, but in Arnhem West, for example, this is not the case as seen in fig.8.

Having visually identified these patterns, we run RFC on these features to make sure whether there is a correlation indeed. That results in an accuracy of about 87% with the following feature importances in fig.9.

If we ignore the train types that we see prevailing here and try classifying only on the capacity of the yard used and the number of trains stopped for a long time, the accuracy drops to 76%. We can argue that is still a good result—it proves there is a pattern there, indeed, since we classify six different locations.

5 Responsible Research

The most critical aspect of this research concerns data access and integrity. ProRail has provided us with sensitive data, and it is essential that this data remains confidential throughout the research project and is not disclosed to any third parties outside the agreement. An NDA (Non-Disclosure Agreement) was signed at the project's beginning, making this confidentiality agreement official. Upon project completion, all provided data will be securely and promptly deleted.

[15] We chose this period because that means the trains are parked even during peak hours

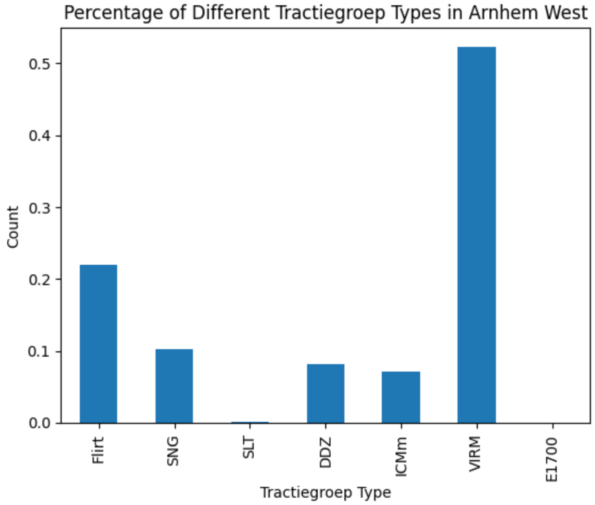

Figure 8: A bar plot showing the distribution of different train types in Arnhem West

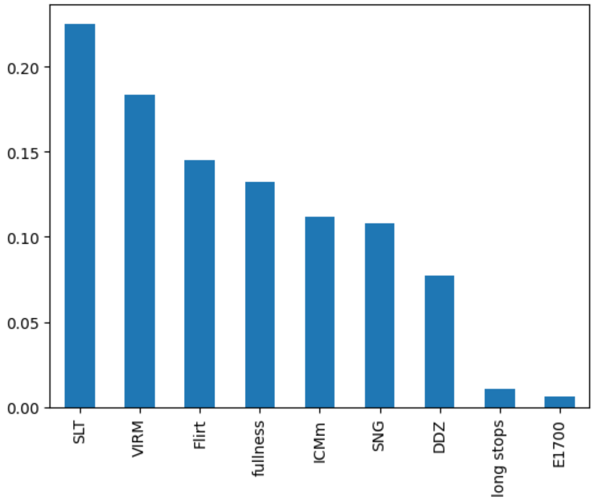

Figure 9: A bar plot showing the feature importance when classifying on train types and capacity features

Regarding data reliability, it is essential to note that ProRail supplied the data which was not collected by our research team. Therefore, we cannot guarantee its reliability or accuracy. However, this does not raise any ethical issues, as the data does not contain elements that could lead to discrimination or bias. Even if there are inaccuracies, they do not introduce any form of unethical bias into the research findings.

The FAIR (Findable, Accessible, Interoperable, Reusable) principles are not directly applicable in this context, as the data used is not publicly available and cannot be accessed without explicit consent from the concerned parties. However, all methods described in the preceding sections can be reproduced, even if the data used is different. Furthermore, all code written in the process of this research is available online. This ensures that the research is transparent and can be validated by other researchers.

6 Discussion

An important point to note is that a successful classification is no proof that the described patterns are the underlying reason for this classification. It only confirms that there are patterns related to those features indeed. Those results can be used to do high-level reasoning about a hypothesis—rule out features we initially thought were useful but do not result in a reasonable classification and include those that do so. In this way, we can narrow down the field that is being investigated.

When considering the train paths as separate entities, we noticed that the lengths of those paths vary across locations, and we can classify on that feature alone. The initial hypothesis was that the differing path lengths were due to the varying sizes of the stations. However, we noticed this pattern even among stations of similar size, indicating that the variation in path lengths is not solely related to the size of the stations. To find the exact reason for that a discussion with experts in the field should be done. Furthermore, in some stations, the entry to the shunting yards appears to take longer paths than other similarly sized ones, as noted in the previous section. This pattern can introduce issues such as delays while entering the yard in order to avoid collisions.

We must point out that because this first approach does not focus on the solutions as a whole, it might miss some correlations in the data. However, that is a good starting point for identifying how each train path is affected in the different locations and giving some intuition about how the whole schedule is influenced.

Investigating the solutions as a whole shows that some locations, like Utrecht, are operating closer to their capacity than others, which could cause future planning issues when the demand increases. At this point, this is not yet concerning as the usage rarely exceeds 70% of the capacity. Furthermore, we saw that some train types are used more often in some stations than others, which could help in planning as planners would stick to similar numbers instead of wasting time investigating something entirely different. The methods mentioned earlier could also be used to keep track of any arising trends and predict whether they could become a problem in the near future. Apart from those findings and geographically specific features, most stations seem to follow similar patterns in their scheduling. That could be explained by the fact that the current solutions are feasible, and most planners stick to the same rules- thus resulting in similar schedules. Moreover, in scenarios where scheduling is involved, consistency is preferred.

An interesting observation is that different layouts do not seem to have that much effect on the scheduling. A possible reason is that FILO (first-in-last-out) setups are easier for planners to work with, even though some trains might block others which could cause additional manoeuvres. Furthermore, we saw that, in reality, not many trains seem to block others judging by the roughly same amount of manoeuvres performed across different locations with different layouts. That seems to be the reason why most stations in the Netherlands use FILO setups indeed. According to experts, apart from easier planning, they require less engineering during the building process when their connection to the main infrastructure is considered.

7 Conclusions and Future Work

In summary, this research focuses on finding train position data patterns that appear across locations. We identify patterns that are generally similar throughout all the locations as well as those that are different. The first ones include how much time trains spend in the shunting yard or the number of manoeuvres they perform inside. The latter features, we conclude, could be useful to automatically classify which location a solution is from. This helps in identifying which locations' specific trends we are interested in are appearing. These patterns include how much of the station's capacity is used and the length of the paths trains take inside the shunting yards.

These patterns could help improve scheduling by providing valuable insights into which parts of the schedules the local search algorithm can add more heuristics to. They can also indicate emerging trends like insufficient infrastructure. Furthermore, these patterns could help us identify which layouts and station designs work better and use that information to improve the future engineering of new locations.

In the future, we could investigate even more patterns in different aspects, which we did not have the time or available data to consider, to find differences/similarities across locations. That could provide an even more comprehensive understanding of the available infrastructure. For example, in which stations occur more delays, which stations are more or less affected during peak hours or seasons. Those are again related to the capacity and possibly the layouts. This would be possible if data from a more extended period were provided or from more locations.

It would be insightful to consider locations in different countries as well since our research is limited to locations in the Netherlands alone. This might mean that country-specific architectures or layouts are missing. Finding best practices in other railway networks that can be applied elsewhere is just as crucial as our findings so far.

References

[1] Roel van den Broek et al. "A Local Search Algorithm for Train Unit Shunting with Service Scheduling". In: *Transportation Science* (2022). URL: https://doi.org/10.1287/trsc.2021.1090.

[2] PI Dalatu and H Midi. "New approaches to normalization techniques to enhance K-means clustering algorithm". In: *Malaysian Journal of Mathematical Sciences* 14.1 (2020), pp. 41–62.

[3] Richard Freling et al. "Shunting of Passenger Train Units in a Railway Station". In: *Transportation Science* (2005). URL: https://doi.org/10.1287/trsc.1030.0076.

[4] Richard Freling et al. "Shunting of passenger train units in a railway station". In: *Transportation Science* 39.2 (2005), pp. 261–272.

[5] Lisan van de Gevel. "How human knowledge can support algorithmic decision-making in the Train Unit Shunting Problem-an exemplary study". In: (2022).

[6] Issa K Hanou, Sebastijan Dumancic, Mathijs de Weerdt, et al. "Increasing the Capacity of Shunting Yards within the Current Infrastructure: A Computational Perspective". In: (2024).

[7] Franck Kamenga et al. "Solution algorithms for the generalized train unit shunting problem". In: *EURO Journal on Transportation and Logistics* 10 (2021), p. 100042. ISSN: 2192-4376. DOI: https://doi.org/10.1016/j.ejtl.2021.100042. URL: https://www.sciencedirect.com/science/article/pii/S2192437621000145.

[8] Vrushali Y Kulkarni and Pradeep K Sinha. "Random forest classifiers: a survey and future research directions". In: *Int. J. Adv. Comput* 36.1 (2013), pp. 1144–1153.

[9] Eduardo José da S Luz et al. "ECG arrhythmia classification based on optimum-path forest". In: *Expert Systems with Applications* 40.9 (2013), pp. 3561–3573.

[10] Evertjan Peer et al. "Shunting Trains with Deep Reinforcement Learning". In: *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2018, pp. 3063–3068. DOI: 10.1109/SMC.2018.00520.

[11] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. "Hyperparameters and tuning strategies for random forest". In: *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* 9.3 (2019), e1301.

[12] Jonathon Shlens. *A Tutorial on Principal Component Analysis*. 2014. arXiv: 1404.1100 [cs.LG].

[13] Kristina P Sinaga and Miin-Shen Yang. "Unsupervised K-means clustering algorithm". In: *IEEE access* 8 (2020), pp. 80716–80727.

[14] Alexander Zien et al. "The Feature Importance Ranking Measure". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Wray Buntine et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 694–709. ISBN: 978-3-642-04174-7.