Master's Thesis

Prediction of energy consumption of rail freight transports using machine learning

T. Straathof (Timo)

August 2022

Supervisors University of Twente

Dr. E. Topan

Dr. I. Seyran Topan

Supervisor LTE Netherlands

Wim Visser

Management summary

Energy prices have increased by a large amount in the last years. This has a large influence on companies that rely on electricity for their operations. LTE Europe is no exception. The company facilitates transportation of goods throughout Europe, primarily by cargo trains. With this large increase, options to decrease the costs of operations such as electricity to power freight trains are examined. In this research, a closer look is taken into the calculation of the used energy on behalf of the Dutch branch of LTE Europe, LTE Netherlands. As the costs for energy is related to how much energy is said to be used, a wrong calculation or estimation can have large effects.

With the use of data available to the company, we have come up with a model that calculates the energy consumption for given transport sections, where a section is a part of the train journey from begin to end where the characteristics, such as composition and locomotive used, are the same. The current method for determining energy consumption is composed in a research into the necessary forces to make a train move and the characteristics of different locomotive types. This results in a formula to estimate the energy consumption, which takes the form of $Energy = Distance * P * Weight^{Q}$, where P and Q are parameters dependent on the type of locomotive and type of grid. The goal of this research is to identify the factors in the data that have influence on the energy consumption of a transport, and to determine whether it is beneficial for LTE Netherlands to move from the current method of estimating the energy consumption to the actual measurements from meters located in the locomotives from an operating cost point of view.

We examine different options for building this model. The decision has been made to use statistical learning methods in order to find a model that would predict the measurements as close as possible. By combining the information available to the company, one comprehensive dataset was created that included information from multiple sources. Information present in this dataset includes factors such as the weight, distance, type of current on the overhead wires, time of day, month, and a calculation of the measured energy use during the time a locomotive is on a certain grid. Although considered, no variables that could be derived from the available data, such as acceleration if the speed of locomotive is available, were added to the dataset.

Two statistical learning methods have been used to create models. These methods are regression, which creates a linear model, and random forest, which results in a black box model that allows nonlinear relations. Both these models are trained on a subset of the dataset, with the complementing set used for testing of the model. This same part is also used on the current estimation of the model to give a fair comparison of the different models. It showed that the regression model consisting of selected single and combined variables performed best on the test data, with the random forest model followed closely. The current estimation model performed the worst of the three models, which indicates that this model is the least accurate.

Additionally to be the least accurate, we found the current estimation model to generally overestimate the actual energy use. As the outcome of this model determines the amount of energy that is paid for, it leads to the conclusion that it is cheaper for LTE to use the metered energy consumption sent from the locomotives to the infranet every 5 minutes for the determination.

The main factors present in the data that influence the energy consumption are the weight, distance, type of locomotive, and type of current on the overhead wire. These factors determine the outcome of the regression model. From these, the weight and distance influence the energy consumption the most.

Upon further analysis of the results of the best performing model, the regression model, we found some inconsistencies that would require further research to improve the prediction power of the model. In cases where the freight load and transport distance are both low, the model output produces unrealistic results. This has two possible explanations. The first is a bias in the training data set, which could have contained too little examples of this particular scenario. The second explanation is the determination of the energy consumption from the measurements. This information is sent out every 5 minutes. It is possible that, for small stretches, a measurement has been allocated to a different part of the transport.

In summary, we find that weight and distance have the most effect on the energy consumption. The models predicting the energy consumption based on the available data showed that they are more accurate than the estimations of the model currently in use for determining the amount of energy to be billed, decreasing the root mean square error of the predictions by 32%. For LTE Netherlands, we recommend to use the meters in the locomotives for determining the amount of energy to be billed instead of the mathematical formula, as the energy consumption predictions resulting from our model are in general lower than those of the current estimation method. There are no implementation limitations as this is already possible within the current system after approval of the meters for this use.

We determine several factors that influence the energy consumption of trains from real world data. Additionally, we have found an improved model for the prediction of energy consumption of freight trains using four characteristics, namely weight, distance, type of locomotive and type of grid. Further research can be used to improve accuracy of the models. An additional factor that has not been included in this research but that is interesting for the operation of LTE Netherlands is the composition of the freight, for example trains moving one type of wagon or one type of material compared to trains moving a plural of these types. External factors from the perspective of the company, for example natural effects such as heavy winds, have also not been discussed in this research, while the possible influence of such factors are of some interest. Finally, train movements not related to transporting freight have been ignored in the scope of this research. It can be worthwhile for LTE Netherlands to get insight in the factors that influence the energy needed for this activity.

Preface

It is a pleasure to present you my master’s thesis. This thesis is the result of many months of labour and concludes my time as a student at the University of Twente. I am grateful for the opportunity granted by LTE Netherlands to have me perform my final internship with them.

Firstly, I would like to thank my supervisor at LTE Netherlands, Wim Visser. The discussions we had were always very pleasant and gave me plenty of insights and ideas that I could use to continue my research. The support I received was very much appreciated. Secondly, I would like to thank my supervisors from the University of Twente, Engin Topan and Ipek Seyran Topan, for the advice, comments and feedback I received on my thesis. I am grateful for the discussions we had, which have helped me stay focussed and motivated all the way through until the end, as well as providing me with knowledge I could use whenever I reached a roadblock.

Finally, I would like to thank my friends and family, for providing mental support and a hearing ear whenever I needed it. I definitely could not have done this without their personal advice, feedback, distractions when needed, and interest in my work.

Timo Straathof

Enschede, August 2022

Contents

1. Introduction

1.1. Company introduction

LTE Europe is a freight company that specializes in the transport of goods by train. The company has offices in 9 European countries, with operations spanning from the Netherlands to Central Europe to Romania. The company works with subsidiaries and affiliates for the transport of the goods as well as their own fleet of locomotives. Additionally, the company offers services for intermodal transportation.

The profit margins in the rail cargo sector are slim. With a sizable number of competitors, it is increasingly hard to compete on price while making a profit. A net margin of 1 to 2 percent on a transport is a good result. As a result, any saving on the costs of the transport can have a large effect on the operational results.

The energy costs of LTE Europe account for 10 to 20 percent of the costs of each transport. Most trains in the fleet of LTE Europe run on electricity. Compared to diesel locomotives, electric locomotives are eco-friendlier as well as more powerful, which are both factors that need to be taken into consideration to be able to operate in this sector. The electricity needed to power these locomotives is purchased in each country separately. The methods to determine how much energy is used can differ. Each locomotive in the fleet of LTE Europe carries a meter that determines how much electricity is used by the locomotives during their stay in a country, both while travelling and while idle. Some countries use the information from the electricity use measured by this meter to determine the amount of electricity for which has to be paid. Other countries use a formula to estimate the electricity used. In the Netherlands, the latter was the situation. However, the purchase of electricity by means of train meters has recently been allowed, provided that the meters has been approved by Erex, the organisation that controls the division of the costs related to the electricity on the overhead lines in the Netherlands.

The purchasing of electricity for rail transport in the Netherlands is arranged by third parties, who charge associated members, including LTE, for their energy use. There are two organisations for the shared net operating under direct current voltage and the high voltage alternating current network. For the direct current network, associated members of the third parties are asked yearly for an estimation of the volume of energy they think they need. Companies pay an amount given this volume at a fixed price at the time of signing the contract. An estimation model determines how much energy is used for a transport. This model takes different factors into account for commuter and freight transports. When the energy used is higher than the estimates of each company, an additional invoice has to be paid based on the share of the volume each company uses. Activities such as shunting and stabling have an influence on the total energy consumption, however these are hard to calculate. For the alternating current net, this calculation is simpler. The total consumption on this network is allocated to each company making use of the net, based on their transports. Due to the lower amount of energy needed for these activities, we will focus on the energy needed for freight transport in this research.

1.2. Research plan

1.2.1. Problem description

LTE wants to lower their operational costs. An opportunity for this came up when in July 2020, a new calculation method for the energy use was introduced. The new method leans a theoretical, physics background. More interestingly, it is now allowed to use the meters in the trains for determining how much energy a train uses, provided that the meters in the train are approved. All the locomotives owned by LTE and their partners are already equipped with a meter. If the energy use measured by the meters is lower than the calculated energy use, LTE should take action to get these meters approved. However, it is uncertain if LTE would benefit from this as it is unknown whether the measured energy use is indeed lower.

In recent times, energy costs have risen a lot. Calculations by the Dutch Central Bureau of Statistics (CBS) estimated an increase of 86 percent of the energy costs for a household (CBS, 2022). With this increase in costs, changing the method of calculating the energy use can have a large impact.

The offered possibility to allow measured energy use also raised questions about what else LTE can do to improve operations. Besides the direct possible savings from the purchase of energy, insight in what factors contribute to the energy use could be used to identify possibilities for energy savings.

One part of the operating costs of LTE are the costs of purchasing energy for the movement of the transports. At the moment, the energy used for the part of the transport within the borders of the Netherlands is determined by a model that estimates the energy consumption. This model has been drafted by an external bureau on behalf of the railway operators active in the Netherlands and uses a theoretical approach, taking characteristics of types of locomotives and calculations of the forces necessary to move a train into account. It is unclear if freight transports executed by LTE perform in accordance with this model or not, therefore creating a possibility of cost saving for the company.

From the problem cluster, the following problem statement can be derived:

Because of a lack of insight in the composition of energy use, LTE faces operational costs that are possibly too high due to inefficient practices and high purchasing costs.

1.2.2. Research objectives

The objective of the research is twofold, as can be seen from the problem cluster (Figure 1). The first objective is to provide insight in the composition of the energy use of trains and what influences this. With this information, new possibilities can be explored to determine whether the energy costs can be lowered. The second objective is to determine whether it would be beneficial for the company to push for their meters to be approved. The top row of the problem cluster shows a problem that is perceived by the company. With the current knowledge of the company, it is unclear whether putting in effort to use the metered energy use will yield lower costs. As the energy consumption is estimated per transport and metered per locomotive, the focus will be on providing insights on a transport level.

Summarized, the objective of the research is:

Find a model that explains the relationship different factors have on the energy consumption of a transport and compare the performance of the resulting model to the current used formula.

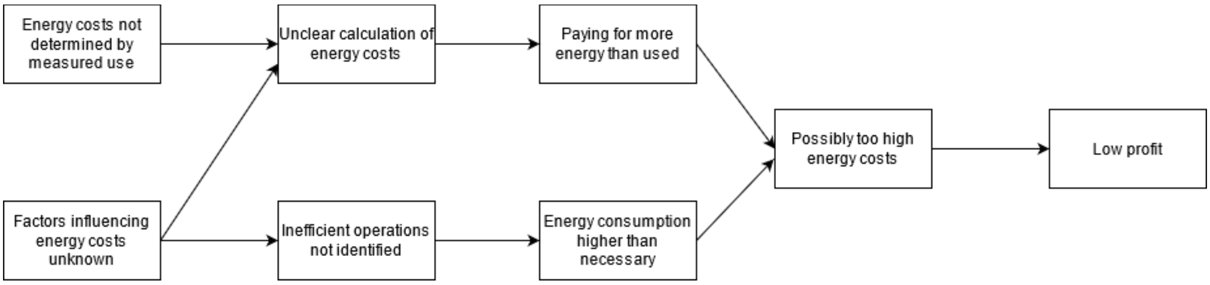

Energy costs not determined by measured use
Unclear calculation of energy costs
Paying for more energy than used
Possibly too high energy costs
Low profit
Factors influencing energy costs unknown
Inefficient operations not identified
Energy consumption higher than necessary

Figure 1: Problem cluster

1.2.3. Research questions

The research questions can be divided in different subjects.

Current process

- What is the method currently in use to determine energy consumption?
- What is the scope of the current method?
- What critique can be given on this method?

Previous research

- What factors are known in literature to influence energy use of railway transport?

Modelling methods

- What methods are available to find an energy use model from available data?

- What metrics can be used for the evaluation of a formulated model and comparison to another model?
- What methods can be used for validation of the model?
- What is the performance of the current estimation model?

Model performance

- What features are important for the prediction of energy use?
- What is the value of the prediction model resulting from the data in comparison to the current estimation model?

1.3. Research approach

The steps taken for this research is based on the Managerial Problem-Solving Method (MPSM) (Heerkens & Van Winden, 2017). The MPSM consists of seven steps:

1. Problem identification (chapter 1)
2. Problem approach (chapter 1)
3. Problem analysis (chapter 2)
4. Formulating solutions (chapter 3)
5. Choosing a solution (chapter 4)
6. Implementing the solution (chapter 5)
7. Evaluating the solution (chapter 6)

In the context of this research, the problem analysis will consists of reviewing the current methods as well as the data availability. Solutions will be formulated by means of a literature review.

2. Context Analysis

2.1. Current estimation methods

The calculation of the energy payments LTE makes in the Netherlands goes through two third parties which have their own method of calculating. These parties divide their areas between the conventional railway net with direct current overhead wires (VIVENS) and the separate alternate current routes (CIEBR). As the current has influence on the energy use, different calculation methods are used. Additionally, there are different calculation methods for rail freight transports and passenger transports. In the continuation of this report only rail freight transports will be taken into account.

A commissioned independent research in 2019 has resulted in formulas for the energy use for both the direct current and the alternating current networks. Using simulations on the routes that are most-used for freight transport, the research found a base formula with parameters based on the type of locomotive and the simulated location. Weighted average parameters for the entire network, weighting the most-used routes heavier, are also given. The formula resulting from this research is given in equation 1:

$$E = Distance * P * Weight^{Q} \qquad (1)$$

where P and Q are fixed values given the type of locomotive and the type of network for the trip. We will analyze the performance of this method in section 4.3.2.

The problem with this method is that it is unclear for LTE if the outcome of this formula aligns with the energy consumption of the trains as measured by the meters in the locomotives. As this information is available to LTE, they want to find out if the energy consumption predictions based on this information result in lower numbers and therefore lower costs, as well as other predictors in the data that have been excluded from this formula.

2.2. Other factors

In addition to the factors included in the formula, there are other things that have influence on the energy use of trains. From the above mentioned report, the calculated energy use can be influenced by a number of things. First, *the train type* can influence it. For freight transport this can mean the type of freight, thus whether the whole transport consists of the same type of wagons or if it consists of multiple types. The second thing is the way a driver drives the *train*. An example of this is whether a train brakes heavily or if it rolls to diminish speed. The third thing are unplanned stops. For each stop, a train has to start moving again, which costs more energy compared to a scenario where the train does not have to make a stop. Finally, a *lower maximum allowed speed* can influence the energy use. For a lower maximum speed, less energy is needed to reach this speed level.

Additionally, there are other factors not mentioned in the report, as the necessary data is either not available or difficult to obtain. Strong winds can disable a train from reaching maximum traction by pushing it to the edge of the rails, increasing friction, and thereby increase the energy needed for moving forward. The altitude of the location where the train is going can also influence the energy consumption as changes in elevation cause an increase in energy consumption when going uphill, or a decrease when going downhill. This factor is omitted for time constraint reasons.

2.3. Data

As the scope of this research is to find a model from data, it is also necessary to explore the available information. In this section the different datasets that are available are explained.

The data available spans 5 months, from January to August 2021, and comes from three sources. This has some challenges which will be elaborated on later in this section. The sources are LTE, Erex, and ÖBB infra. Each source collects different information. In this section, the information from each source will be explained. Additionally, information that can be derived from each source will be highlighted, followed by how each source relates to one another. Finally, difficulties for the research will be identified.

2.3.1. LTE dataset

The first data source is LTE. LTE themselves keep track of the transports they execute. Information in this database has that focus and is not used to keep up with the locomotives in real time. Each entry covers one transport. A transport means a locomotive with carts that travel with the goal of either transporting goods or transporting empty carts to the requested place. The information entails among others a departure and arrival time and date, the destination, amount of carts involved in the transportation and much more. Some of the input in this dataset is non-numerical and contains information for planners.

2.3.2. Erex

The second source of data is the Erex system. Erex is a third-party system used by the association of railway transport companies in the Netherlands. As mentioned in the introduction, the payment of electricity for the locomotives is handled via third parties. To determine the amount that needs to be paid, the necessary calculations are made based on the information stored in Erex. The information is partly provided by the companies that make use of the railway network, partly metered, and the outcome of the calculations is also obtainable.

The data available in the Erex system is extensive and covers the *weight of trains*, the *location* at a time of day, *energy consumption* and plenty more. The information that is available and which can be used consists of train run data. A train run is a part of the transport that takes place under the same circumstances, which means that no changes in the train composition occur in this period. Similarly, the catenary voltage in a train run is the same as well. A transport can therefore consist of multiple train runs.

Each entry in the data gathered from Erex entails a train part run. The entities in the train run dataset are the following:

- *Train ID* – A unique ID for a train.
- *Operating day*
- *Part start* – The start time of the train run
- *Part end* – The end time of the train run
- *Traffic category* – Type of transport (e.g. goods)
- *Traction unit set* – A unique identifier for a traction unit
- *Price category* – Either peak or off-peak. This is used for the payment of energy used as energy during peak hours is more expensive.
- *Grid* – Location of the train run (e.g. conventional rail network, Betuweroute)
- *Calculation type* – Method of energy use calculation (e.g. estimated, metered)
- *Estimated consumption (kWh)* – Estimated consumption of the transport during the train run part
- *Metered consumption (kWh)* – Metered consumption of the transport during the train run part
- *Metered generation (kWh)* – Metered generation of the transport during the train run part
- *Net consumption (kWh)* – Net consumption during the train run part
- *Weight (tonnes)*
- *Distance (km)*
- *Start coordinate (latitude* and *longitude)* – start position of the train part run (e.g. departure place of the train or place where the train switches from the conventional rail network to the Betuweroute)
- *End coordinate (latitude* and *longitude)* – end position of the train part run (e.g. destination)
- *Train part result type* – Calculation type used for payment

In this dataset, the energy consumption is depicted in either estimated and metered consumption or net consumption. When the calculation type is estimated, the energy consumption will only show up under net consumption. If the energy consumption can be metered, this will show up in the metered consumption. The estimated consumption based on the formula for the given characteristics (grid, weight, distance) is given in that column. Additionally, locomotives can transfer some energy back to the overhead wire. This generated amount is tracked under *metered generation* and subtracted from the *metered consumption* to give the *net consumption*. The *net consumption* is the variable that we want to predict.

2.3.3. ÖBB infra data

The final source of data is ÖBB infra. ÖBB is the Austrian federal railway. As the headquarters of LTE Europe are located in Austria, the meters in the locomotives are connected to their system. The data here is sent from the locomotives. The entities are:

- *Locomotive ID* – The number on the locomotive
- *Latitude* – Coordinates of the location of the train at the time of sending the data
- *Longitude* – Coordinates of the location of the train at the time of sending the data
- *Average speed* – Average speed during the 5 minutes since last sending data
- *kWh+* - Energy consumption during the 5 minutes since last sending data
- *kWh-* - Energy generation during the 5 minutes since last sending data
- *kvarh+*
- *kvarh-*
- *Country* – 2-digit code representing the country where the locomotive is at time of sending the data

This information is sent out every 5 minutes. The kvarh parameter is nonzero whenever the locomotive is connected to an alternate current (AC) catenary. The country parameter is followed as the locomotives of LTE Europe cross borders. Together with the latitude and longitude coordinates, the location of the border crossing can be determined. This can be depicted in the ÖBB infra system on a map.

2.3.4. ProRail

The final data source is the information as provided by ProRail. The information here is used as input for the information in Erex and is partly provided by LTE (weight) and partly collected or assigned (train number, travelled routes, actual departure times).

2.3.5. Dataset conclusion

In total, from these different sources we have information about 1,337 transports, cut up in 6,855 train run parts, and information from 17 locomotives, resulting in 527,650 points in time with information.

2.3.6. Challenges

The timeframe of all collected data is between January and August 2021. However, the LTE data is transport oriented; it gives information about the transport from a business perspective. The Erex data on the other hand is energy oriented; it keeps track of the (estimated) energy and energy costs of each transport by following the routes. Then, the ÖBB infra data is locomotive oriented; the data is sent from each locomotive from the moment it connects to a catenary to the moment it shuts down. From the earlier mentioned entities, one main challenge can be identified:

- *The timeframe of the sources is different*

This problem mainly arises with the ÖBB infra data. All information is given in 5 minute intervals. In general, the energy use per transport is wanted.

3. Literature review

3.1. Energy consumption models

Research in the area of energy consumption of vehicles have focused mainly on electric vehicles. Prediction models based on real world data and physics approaches are plenty available. This is however not the case for rolling stock. Most energy consumption models for trains are centred around passenger railway systems and have mainly been approached from a physics point of view. The resulting models are generally a backwards simulation of the power flow, looking from the vehicle dynamics back to the power unit. As traction energy is the largest factor in both passenger and freight railway (Su et al., 2016), methods for modelling the energy consumption for passenger railway can tell us relevant factors. However, traction energy does not account for the entirety of the energy consumption as energy here is also used for heating compartments.

Yun et al. (2009) define traction energy as the energy used to overcome the work of resistance of the train, increase the kinetic energy to compensate for the kinetic energy loss resulting from braking, and supply for the gravitational potential difference. The total energy consumption depends on the traction and the energy efficiency. Su et al. (2016) use a model that includes the train speed, the maximum traction force and the relative traction force over the trip time for the Beijing metro system to find an optimal train control model. Additionally, the gradient of the track, the weight of the train, and the traction and braking force of the traction unit influence the energy consumption related to traction (Nespoulous et al., 2021; Wang & Rakha, 2017). External factors such as wind and contact forces also influence the energy consumption (Nespoulous et al., 2021).

3.2. Machine learning

Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It focuses on the development of computer programs that can access data and use it to learn for themselves. Machine learning enables analysis of massive quantities of data and while it generally delivers faster, more accurate results, it may also require additional time and resource for proper training.

Different steps of the machine learning process will be explored in the following sections.

3.2.1. Characteristics of machine learning

Four machine learning types are:

- *Supervised* machine learning, which algorithms can apply what has been learned in the past to new data to predict future events. It can also compare output with the correct, intended output and find errors in order to modify the model accordingly.
- *Unsupervised* machine learning algorithms are used when the information to train is neither classified nor labelled. It studies how systems can infer a function to describe a hidden structure from unlabelled data. It does not figure out the right output.
- *Semi-supervised* machine learning algorithms use both labelled and unlabelled data for training. This is chosen when the acquired labelled data requires skilled and relevant resources in order to train it.
- *Reinforcement* machine learning algorithms interact with its environment by producing actions and discovering errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. It allows for determination of ideal behaviour within a specific context (*What is Machine Learning? A Definition.*, 2020)

To simulate unseen data for the trained model, available data is subjected to *data splitting* whereby it is split to 2 portions (train-test split). The training set is the larger set, accounting for 80% of the original data. The testing set accounts for the remaining 20%. The model trained by the training set is applied on the testing set to make predictions. Selection of the best model is made based on the performance of the model's performance on the testing set. Hyperparameter optimization may also be performed to obtain the best possible model.

Another method splits the data in a *training*, *validation*, and *testing* set. The validation set is used for evaluating the predictive model whereby predictions are made, model tuning can be made (hyperparameter optimization for example) and the best performing model can be selected (Nantasenamat, 2020).

3.2.2. Data preprocessing

Data pre-processing is the process by which the data is subjected to various checks and scrutiny in order to remedy issues of missing values, spelling errors, normalizing/standardizing values such that they are comparable, transforming data, et cetera. The quality of data is going to exert a big impact on the quality of the generated model. To achieve highest model quality, significant effort should be spent in the data pre-processing phase. (Nantasenamat, 2020).

Additionally, *exploratory data analysis* can be performed to gain a preliminary understanding while getting acquainted with the dataset. Three approaches are descriptive statistics (mean, median, mode, standard deviation), data visualisations, and data shaping (pivoting/grouping/filtering data).

3.2.3. Modelling techniques

There are several modelling techniques within machine learning. In this section we discuss these methods and look at the situations in which each method can be applied. As the two main reasons to find a function are prediction and inference, there is a trade-off between flexibility and interpretability of a model. Flexibility leads to better predictions, but the effect of the features on the outcome is less clear. A model that is easy to interpret has generally worse predictions.

Linear regression is a very simple approach for supervised learning. It can be used to evaluate a range of possible formulas, which can include multiple variables as well as higher dimensions (polynomial regression). The downside of polynomial regression is that it allows wild behavior at the tail ends due to its global nature. Solutions to combat this behavior are regression splines and smoothing splines. Assessing the accuracy of the model can be done by means of the R^2 value and the residual standard error.

Resampling methods involve fitting the same statistical method multiple times using different subsets of the training data. Two of the most commonly used resampling methods are *cross-validation* and the *bootstrap*. In cross-validation (CV), a subset of the data is set aside and used for testing the model that results from the training data. It can be used to determine how well a single statistical learning method is performing with different flexibilities, for comparing different learning methods, or to determine how well a given learning procedure can be expected to perform. The MSE is used as an indication of validation set error. Alternative cross-validation methods are of the *k-fold cross-validation* form, where the data is divided in *k* folds. Each fold of data is used once as a validation set and part of the training set when not. The estimate of the CV is the average of the validation set errors. Bootstrapping is a resampling method where *n* observations from the data set are selected with replacement to produce a bootstrap data set. It can be used to quantify the uncertainty associated with a given estimator or statistical learning method. It can be applied to a wide range of learning methods.

Support vector machine is an approach that lends itself best for classification. A hyperplane in p-1 dimensions, where p is the amount of variables of the observations, separates the data points into two classes. The distance from an observation to the hyperplane, also called *margin*, is maximized to get the maximal margin hyperplane. Support vectors are observations close to the hyperplane that influence the position of the hyperplane. In general these are linear. For non-linear separation, a (higher-dimension) kernel function can be used. Support vector machines are computationally efficient and robust against overfitting, but are difficult to interpret.

Tree-based methods segment the predictor space into a number of simpler regions. These methods can be used for both classification and regression. Tree-based methods are easy to interpret but lack prediction accuracy. By means of *bagging*, *random forests* and *boosting*, prediction accuracy can improve but interpretability will decrease.

A tree is created by binary splitting the data into two smaller regions, until each region has fewer than a set number of observations. The predictor and cut point at each step are selected such that the resulting tree has the highest accuracy. Each observation in a given region has the same predicted outcome or class. This method may lead to overfitting the data, but can be combatted by pruning the tree, where a trade-off is made between the complexity of the tree and the fit to the training data.

Decision trees suffer from high variance. *Random forests* are a method to reduce the variance. A number of decision trees are assembled from a subset of data. This training data is bootstrapped from the observations. The difference between *bagging* and *random forests* is the amount of predictors considered at a split. Where bagging considers all predictors, therefore resulting in possibly correlated trees. Random forests prevent this by considering only a subset of the predictors, thereby decorrelating the trees.

Another approach to build decision trees is by *boosting*. Boosting builds decision trees sequentially from a subset of data, where the training data is a modified version of the original data set using information from previously grown trees. Specifically, the insignificant part of the built trees is considered. This method often results in smaller trees, which improves interpretability but has a higher variance compared to random forests. (James et al., 2013).

3.2.4. Feature selection

Feature selection is applied to reduce the number of features in many applications where data has hundreds or thousands of features. Different feature selection methods can be broadly categorized in to the *wrapper* model and the *filter* model. The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets. The filter model separates feature selection from classifier learning and selects features subsets that are independent of any learning algorithm.

Features in an original set can be divided into four groups: completely irrelevant and noisy features; weakly relevant and redundant features; weakly relevant and non-redundant feature; and strongly relevant features. Supervised feature selection should include the latter two groups. To achieve high efficiency, heuristically decide if a feature is relevant if it is highly correlated with the class. Selected features are subject to following redundancy analysis (Yu & Liu, 2004).

Wrapper methods conduct a search of the predictors to determine which, when entered into the model, produce the best results. Many models are evaluated at each step in the determination of the best model. Examples of wrapper methods are forward selection, backward selection, stepwise selection, simulated annealing and genetic algorithms.

Forward selection is a method that evaluates each predictor for inclusion in the current model. At each step, a new model is created for each not included predictor which includes said predictor. Afterwards, the statistical significance of the new model is evaluated. If the smallest p-value is less than the inclusion threshold, the model is updated with the predictor that is the most statistically significant. This repeats until no statistically significant predictors are left out of the model. *Stepwise selection* is a modification of the forward selection where, after each candidate variable is added to the model, each term is reevaluated for removal from the model. This procedure makes the forward selection procedure less greedy.

Backward selection is a similar method where from a starting model which consists of all available predictors, predictors that are not significantly attributing to the model are iteratively removed. *Recursive feature elimination* is a backward selection algorithm that avoids refitting many models at each step of the search by calculating predictor importance. Each evaluated model consists of a subset of the most important variables. The best performing model is selected for the next iteration of the algorithm.

Filter methods evaluate predictors prior to training the model and, based on the evaluation, a subset of predictors are included in the model. Each predictor is evaluated independently from other predictors. As a result, significant but redundant predictors can be selected. Examples of metrics that can be used for this are ANOVA for numerical predictors and Fisher's exact test for categorical predictors. MIC and the Relief algorithm are generic methods for quantifying predictor importance (Kuhn & Johnson, 2013).

Shrinkage methods or regularization are another option to improve the fit of a model by limiting the effect of features on the model. Where feature selection methods limit the amount of features selected for a model, shrinkage methods include all predictors but shrink the coefficient estimates of insignificant predictors. This deliberately increases the bias of the model to reduce variance and improve model performance. It can also prevent overfitting. The most well-known techniques are *ridge regression* and *lasso*. Ridge regression penalizes the squared coefficients. Lasso penalizes the absolute value of the coefficients. This results in the possibility to have coefficients set to zero in lasso, whereas ridge regression will always include all of the variables in the model (James et al., 2013).

3.2.5. Model evaluation

Evaluation of machine learning models is a crucial step before application, as it is essential to assess how good a model will behave for every single case. Many machine learning models are good in overall results but have a bad distribution/assessment of the error.

Common evaluation methods for regression are the mean squared error (MSE) or the mean absolute error (MAE). With a same result for a quality metric, two different models might have a different error distribution (Bella et al., 2010).

Evaluation of the performance of regression models are performed to assess the degree at which a fitted model can accurately predict the values of input data (Nantasenamat, 2020).

- A common metric for evaluating the performance of regression models is the coefficient of determination (R^2): $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$, which is 1 minus the ratio of the residual sum of squares to that of the total sum of squares.
- The mean squared error (MSE) and root mean squared error (RMSE): $MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$

3.3. Conclusion

There is little information about the factors that influence energy use of freight trains. Literature finds several factors that are not easily measured and unavailable in the data we have at hand. Much of this is complicated to measure and is unavailable for use within the scope of this research.

Several methods of machine learning have been discussed. For this research, we will use regression and random forest as techniques. By using two techniques, weaknesses of either technique can be combatted by the other technique, resulting in two models that minimize prediction errors in different ways. This allows for choice in the best suited model for this problem.

We will process the data in chapter 4, and perform explanatory data analysis. The model training will take place in chapter 5, with a comparison between different models in section 5.3. We will use the MSE and RMSE for this comparison as this is the easiest method for comparing models of different learning techniques.

4. Data analysis

4.1. Example train

To get an idea how energy consumption fluctuates over time, we take a closer look at an individual transport. The selected transport follows a route that is frequently used from the port of Rotterdam towards the German border near Bad Bentheim. The particular transport occurred on July 22nd, starting at 22:20 and ending at 04:05 on July 23rd where it crosses the border. In this period, we can identify several separate periods:

i) The startup phase. In this period, the locomotive has been switched on but the transport has not started yet. Shunting also happens in this period

ii) Transport over the harbor line of the Betuweroute

iii) Transport over conventional net, connecting the harbor line and the main line of the Betuweroute

iv) Standstill near Kijfhoek

v) Transport over the main line of the Betuweroute

vi) Transport over conventional net between Betuweroute and the German border

By identifying these different parts, we can examine if things stand out at certain points and look for explanations connected to the activities. The information we compare are the net energy consumption, average speed, and the change in average speed. The information is gathered in 5 minute intervals and taken from the ÖBB data.

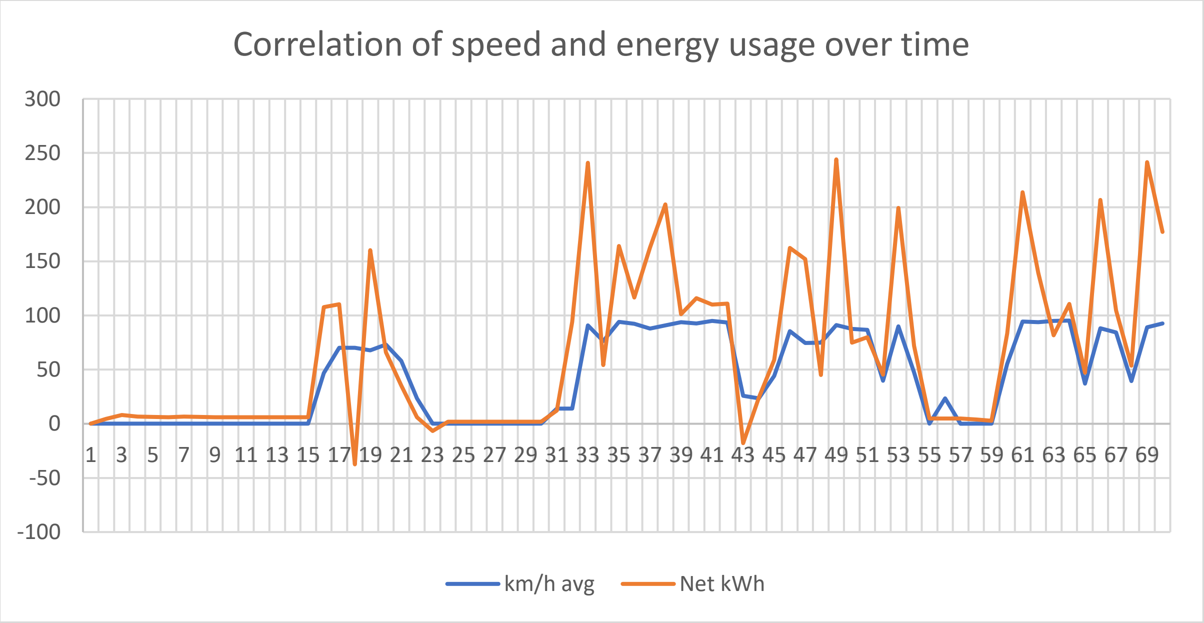


Figure 2: Speed and energy use over time of example train

From Figure 2 we see that the train is idle for 70 minutes before it starts moving. Before beginning the transport, the locomotive extracts 86.5 kWh which is used for starting up the locomotive and whatever shunting is needed. This is extracted from the alternating current grid present in the harbor of Rotterdam. This is lower than the average energy consumption in a 5 minute interval while this locomotive is in motion, which is 107 kWh.

The energy consumption in the phases as mentioned compared to the change in speed between the intervals shows that during startup more energy is used compared to standing still at the conventional net near Kijfhoek, where little to no energy is used, as can be seen in the left of Figure 3. Furthermore, the energy consumption on the Betuweroute (alternating current grid or AC) is lower than on the conventional (direct current or DC) grid. However, the speed on the Betuweroute is much more constant which likely has a positive influence. In general, peaks in energy consumption can be found when the locomotive is accelerating, but not limited to just those periods.

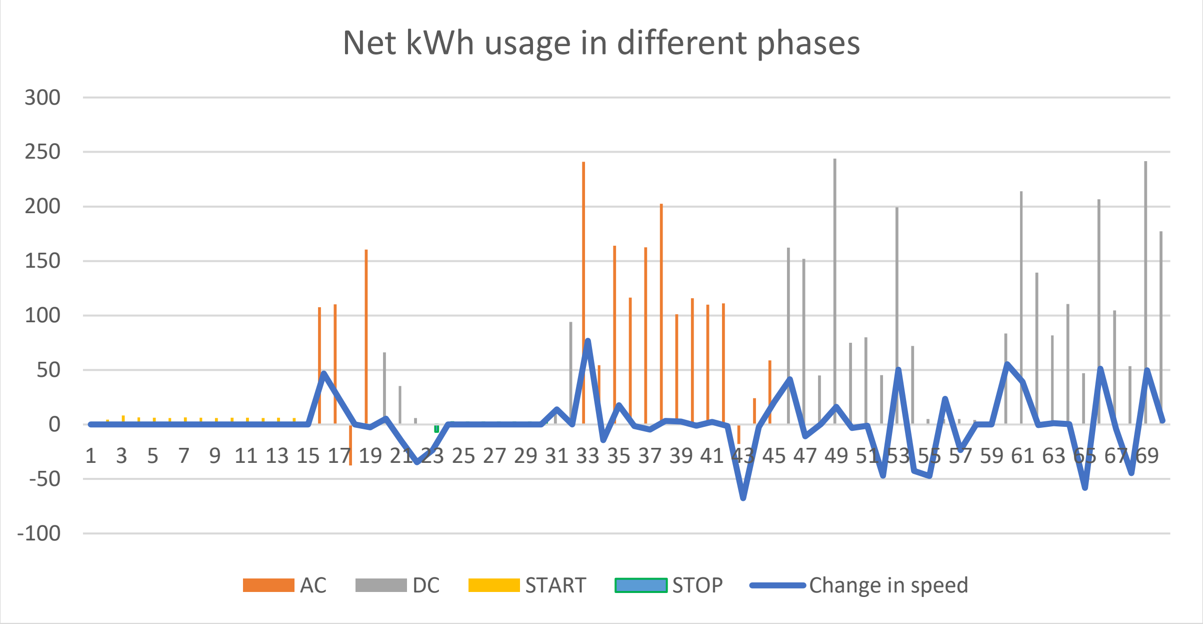


Figure 3: Net kWh usage in different phases

4.2. Data processing

The data as described in chapter 2 requires transformation before it can be used in a regression or machine learning model. Each dataset has its own format, which does not align with the formats of the other sources. The steps taken to get a finite dataset will be highlighted in this section.

First, we select which sources to include in the model. As mentioned in section 2.3.1 and 2.3.4, the LTE dataset is used for the ProRail data. Additionally, the LTE data is gathered with a focus for operational users. The measured information is passed on to ProRail. Therefore, for creating the dataset that will be used for the model, we incorporate only the information from ProRail.

As the scope of the research is to find a model that explains the factors that influence the energy use as well as evaluate the resulting model against the current formula in use, we decide to format the finalized dataset in a similar format. We divide transports into train runs as they are put in the data pulled from Erex. The separating factors for different inputs are i) the start of a transport; ii) the end of a transport; and iii) a change of network (e.g. from alternating current to direct current or vice versa).

The data from ProRail and Erex are combined on the unique combination of train identification number and traffic date. The resulting dataframe is condensed to remove columns that do not present categorical or numerical information. We add separate variables that describe the time of day in which a transport started and the month of the traffic date. Any seasonal factors can be determined if they appear within the available data. The combined database consists of approximately 2,700 entries, or transport parts.

The locomotive data pulled from ÖBB infranet has been filtered to only include entries sent from the Netherlands. The ÖBB data consists of every entry sent by the locomotive when it is in the Netherlands, which are approximately 80,000 entries for 17 locomotives. These entries also include time periods where the locomotive is either idle or preparing for transport, also called stabling and shunting. This data is divided in two parts. The measured energy consumption of a locomotive that is connected to a transport part is added up over all 5-minute entries that fall within the start and end time of the transport part and added to the combined training dataset. All the data entries that cannot be assigned to a transport based on the above mentioned criteria are put together in a second dataset which will be used to analyze the energy consumption during these activities.

Outliers are identified by means of the rejection criteria as set by Erex. These criteria deem all measurements that are less than 50% or over 250% of the estimate of the energy consumption invalid. For these measurements, the estimation will be taken for the payment. Combined with the possibility for erroneous values used in the dataset due to the interval in which information is sent out, these percentages are deemed to be valid to be used as cut-off points for outliers.

4.3. Preliminary analysis

After creating the training dataset, we perform some preliminary statistics and visualizations to see how the data behaves. This is two parted. First, we explore how the measured energy consumption behaves compared to certain variables. This allows us to see whether unusual things occur. Second, we can compare the measured energy consumption against the predicted energy consumption from the formula mentioned in section 2.1. This comparison allows us to get a feeling on whether the estimations from the formula exceed or undermine the measurements.

4.3.1. Visualisations

To get a first feeling what could influence the energy consumption, we compare several variables to the measured energy consumption. We also compare some combined factors, specifically categorical variables with another categorical variable or a continuous variable (e.g. “Network type” with “Distance” or “Time of Day”.

4.3.1.1. Distance

The first effect we analyse is that of distance on the measured energy consumption. Logically, to move a further distance would take more energy. It is therefore interesting to see that there appears a decrease in the trendline of energy consumption in the plot in **figure X** between a distance of approximately 150 and 180 km. Additionally, we see that there are certain distances that have a lot of measurements. This shows some routes that are used more often than not. Given that little to no rail freight transport occurs solely within the Dutch borders, there are certain corridors to (mainly) the German border that are travelled. Therefore, these distances coincide with the length of certain train parts. For example, the maximum distance on the Betuweroute-havenspoorlijn (the port of Rotterdam) is fixed. Similarly, there are certain “exits” a freight train on the Betuweroute-A15 (main track between Rotterdam and the border with Germany) can take depending on the destination.

The decrease is unexpected with the notion that further distances account for higher energy consumption. This would indicate that other factors influence the energy consumption in addition to distance.

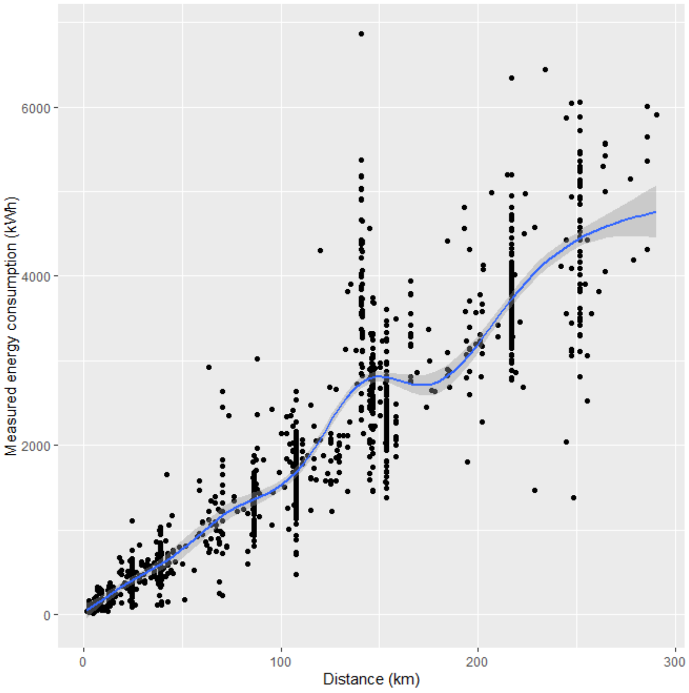


Figure 4: Relation between the measured energy consumption and transport distance

4.3.1.2. Weight

The weight compared to the measured energy consumption shows a less clear correlation as the distance. This can be explained by separate entries in the dataset of the same transport. As these are cut in train parts where the characteristics stay the same, weight is often the one characteristic that stays the same. Therefore, the same transport can show up multiple times in the plot, at completely separate heights.

Several clusters are present. Three of these clusters occur in the middleweight category between approximately 1250 and 1750 tonnes. The first one concentrates between approximately 0 and 500 kWh. The second cluster consists of measurements between approximately 1000 and 2000 kWh. The third cluster contains measurements between 3000 and 4000 kWh. Additionally, there are two smaller clusters for heavyweight transports. These are spread out between low energy consumption (0-1000 kWh) and high energy consumption (5000-6000 kWh). The reasons for these spreads cannot be determined from this plot alone. Finally, there appears a clear correlation between weight and the energy consumption when looking at weight in isolation.

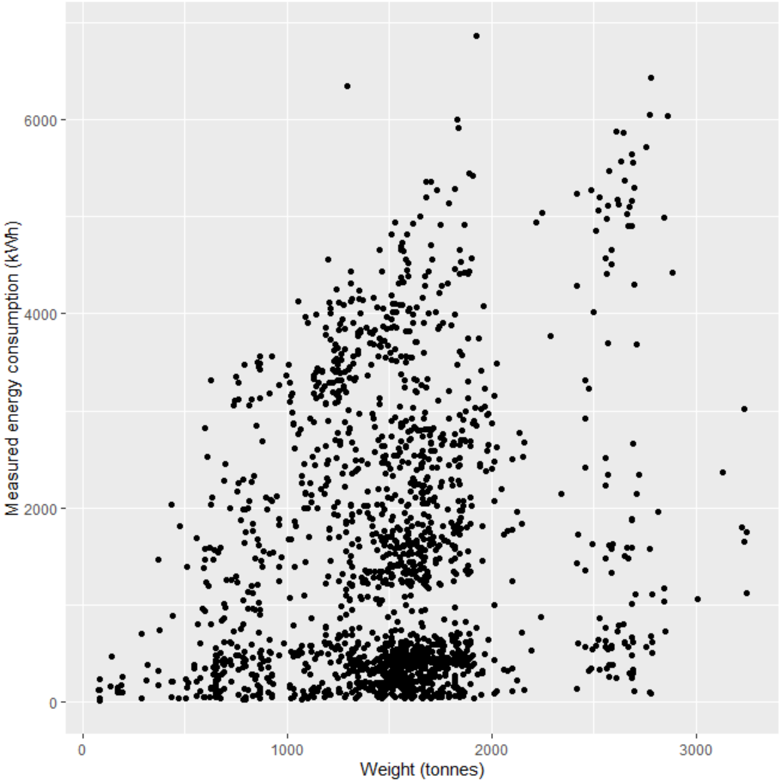


Figure 5: Relation between measured energy consumption and transport weight

4.3.1.3. Network

The freight transports occur over both the 25 kV alternating current (AC) and the 1500 V direct current (DC) network. Given that both networks have different parties that handle the payment of energy, it is interesting to see what the effects on the energy use the different voltages have.

It shows that the energy consumption of transport parts on the AC networks (harbor line and A15 line) is lower than on the direct current network. This could partly be explained by the shorter distances of the AC network, however it seems unlikely that all the difference comes from that. In section 4.3.1.7 we will examine closer what the effect of distance is on the energy consumption.

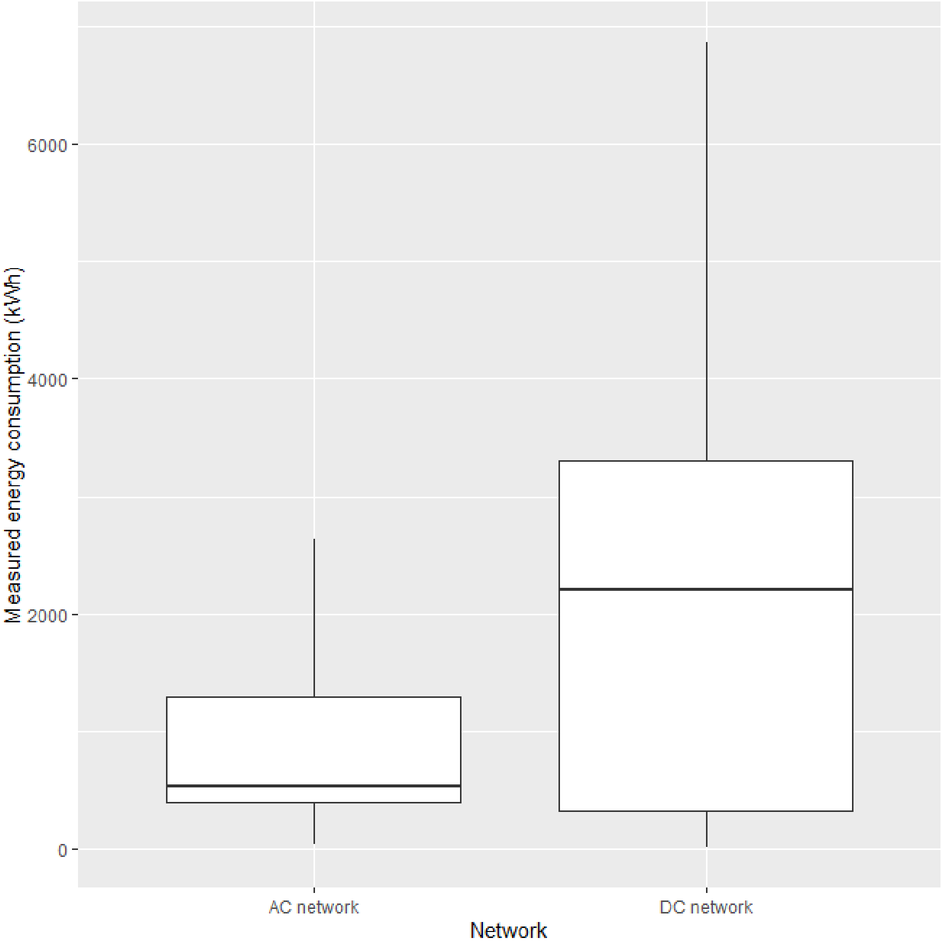


Figure 6: Boxplot distribution of measured energy consumption on different network types

4.3.1.4. Locomotive type

LTE has two types of electric locomotives which they use for their transports, of which data is available. These types of locomotives, also called ‘Baureihes’, have different characteristics in terms of their working (e.g. different brake power, difference in engine power). This can have effects on the energy consumption.

Looking at the locomotive type in isolation, it is unknown whether one is more often used for certain transports that require more energy than the other. We observe that locomotives of type ‘BR193’ has generally lower energy consumptions compared to transports executed by locomotives of type ‘BR186’. This would indicate that the type of locomotive has some influence on the energy consumption.

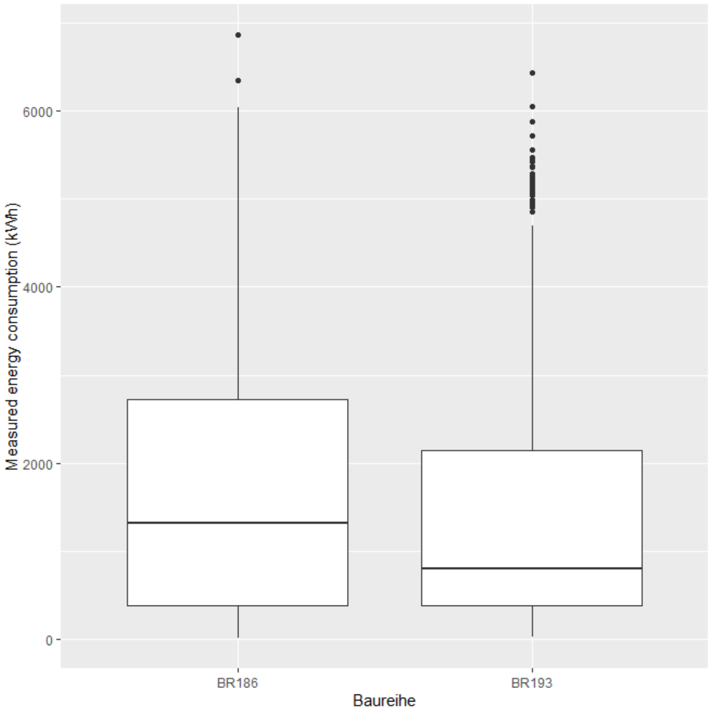


Figure 7: Boxplot distribution of measured energy consumption and locomotive types

4.3.1.5. Months

It is hard to identify external causes of energy consumption, for which no data is available. One way to possibly factor it in is by taking a look at the energy consumption during the months of the data. The underlying assumption for any conclusion that can be made from this reasoning is that the type of transports are equal each month, thereby having external factors be a large factor in the discrepancy of energy consumption during the months. Although this is unlikely in reality due to the nature of demand for transport changing, visualizing the energy consumption for each month could show unusual things.

The energy consumption in the first four months seems to be higher compared to those in the latter four months. This could indicate that there is a seasonal effect taking place. As we have seen in the previous sections, locomotive type and network also appear to have an effect. If a beneficial combination of those was used more often in these months, this could explain differences between the months. The effect this has will be explored when training the model.

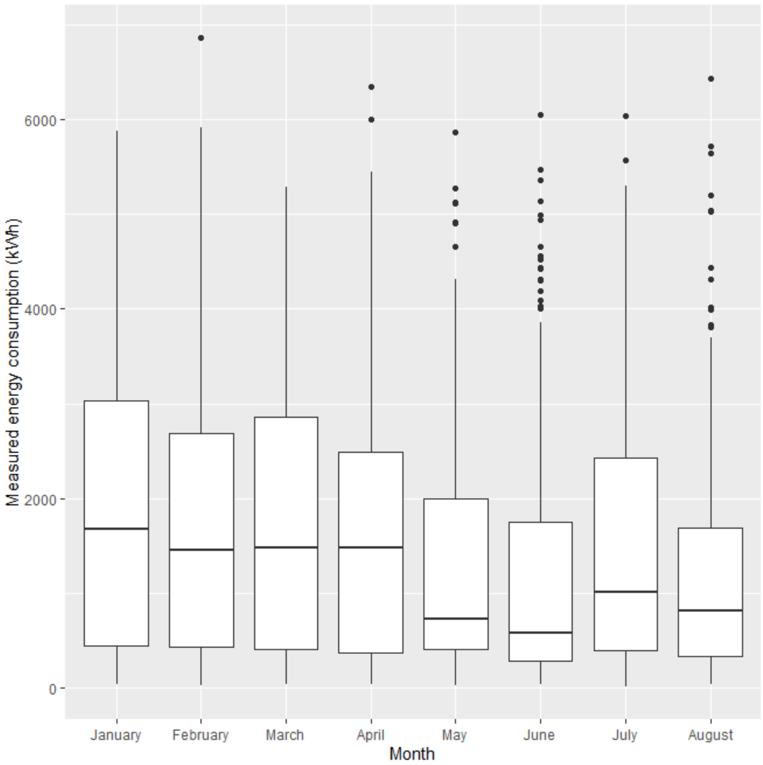


Figure 8: Boxplot distribution of measured energy consumption over different months

4.3.1.6. Time of day

It is known that frequent acceleration and deceleration influences the energy consumption of a locomotive in a negative way. It is very much preferred to keep a non-fluctuating speed. It is difficult to determine the acceleration and deceleration of a train due to the format in which data relating to the speed of a train is sent by the locomotive to the infranet. We try to solve this issue by looking at the time of day. The time of day can indicate heavy traffic, as during the day there is more interference from passenger trains, especially on the DC network, resulting in more decelerations and accelerations.

The tail of evening transports shows considerable lower energy consumptions compared to other times in the day.

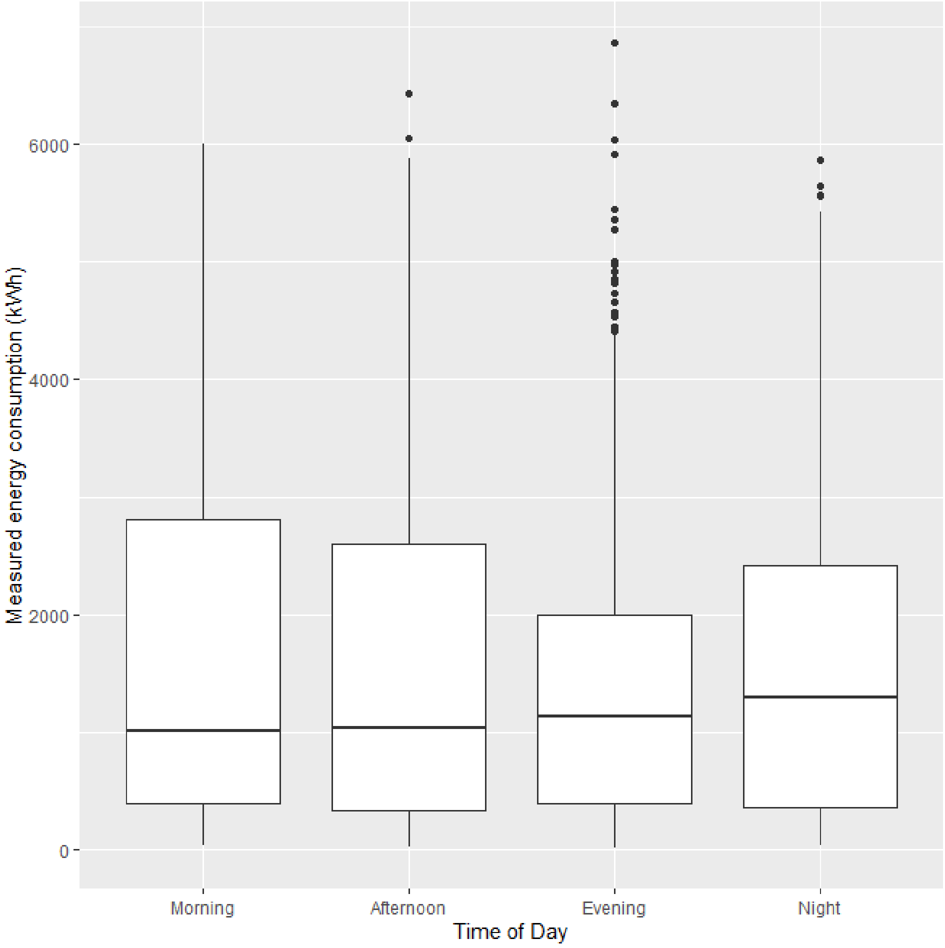


Figure 9: Boxplot distribution of measured energy consumption over times of day

4.3.1.7. Combined variables

We have seen in the past sections some trends when plotting the measured energy consumption against some variables. This does not take interaction effects. To get an insight in the possible presence of these effects, we take a closer look at the effects of distance combined with type of locomotive and distance with type of network, as well as type of network and type of locomotive. We do not use weight for this visual examination as it showed that the correlation between weight and the measured energy consumption is difficult to identify without diving deeper into the data. Plotting weight given other factors does not give clarity into underlying interaction effects. Additionally, there are only two types of network and two types of locomotive, making these variables easy to use for a visual examination as the amount of graphs will be manageable.

In Figure 10 we have plotted the effect of type of locomotive on the energy consumption over an increasing distance. On the left side, we observe a curve in the trendline after 100 km. This can be accounted to other variables or to characteristics of the locomotive. The trendline on the right side for locomotives of type BR193 appears to be more straight in comparison. Between the two sides, small differences can be observed. It appears that the energy consumption for transports executed by locomotives of type BR193 is slightly lower over middle and long range distances, albeit a small difference.

Continuing, we observe that on the AC network only smaller distances are travelled. Additionally, it appears that the energy consumption on this network is slightly below that when travelling a similar distance on the DC network. However, this comparison is hard to make based on this visual (Figure 11).

Finally, we can see from Figure 12 that the measured energy consumption on the AC network is much lower compared to the DC network, likely to do with the shorter distances travelled as discussed above. Additionally, we can conclude that for the DC network the measured energy consumptions are lower for locomotives of type BR193. This can be explained by the locomotive characteristics, but can also have to do with the distribution of trips over the different locomotive types.

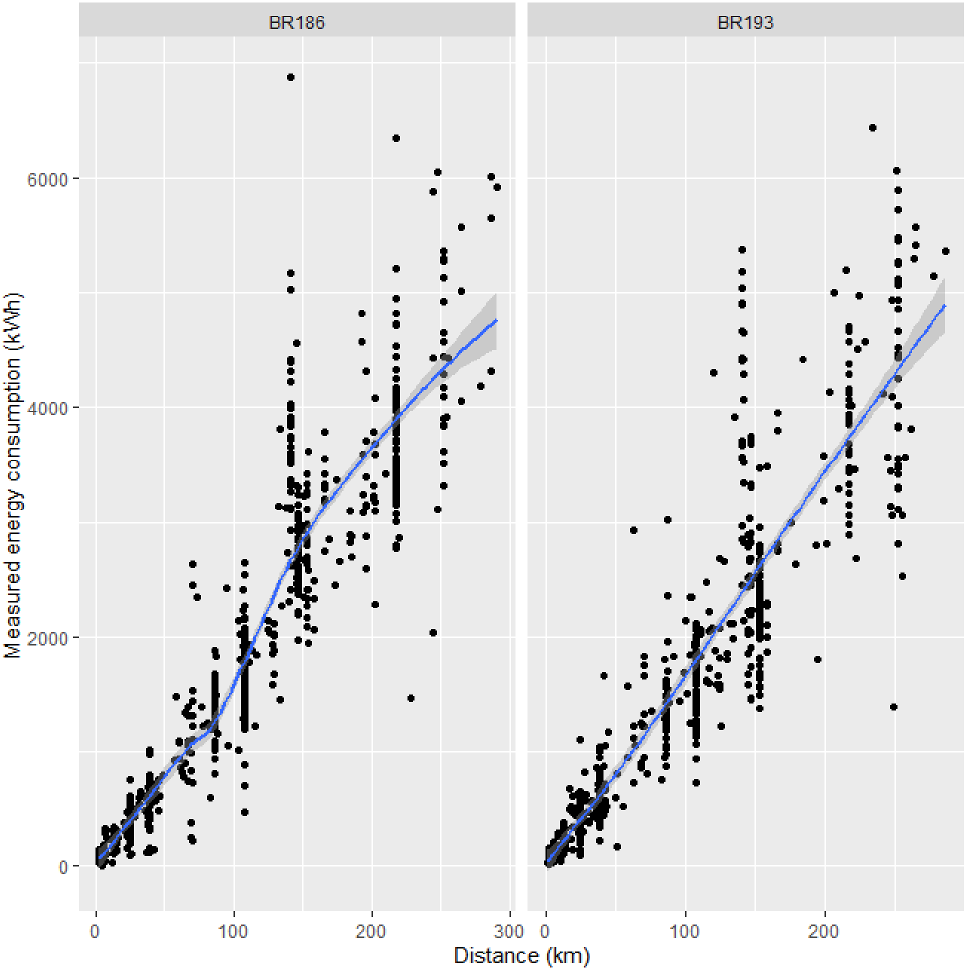


Figure 10: Correlation between measured energy consumption and transport distance given type of locomotive

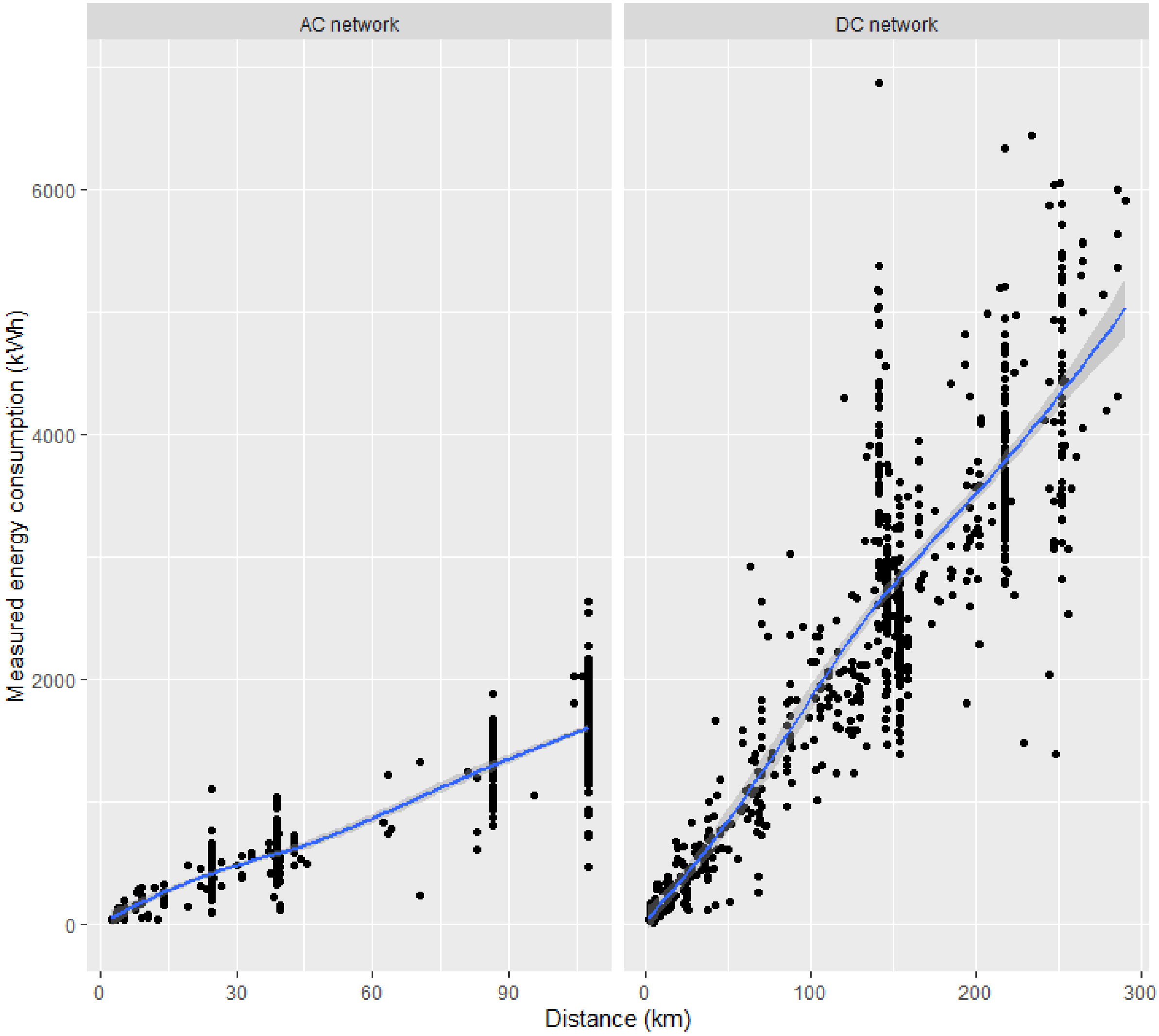


Figure 11: Correlation of measured energy consumption and transport distance given a type of network

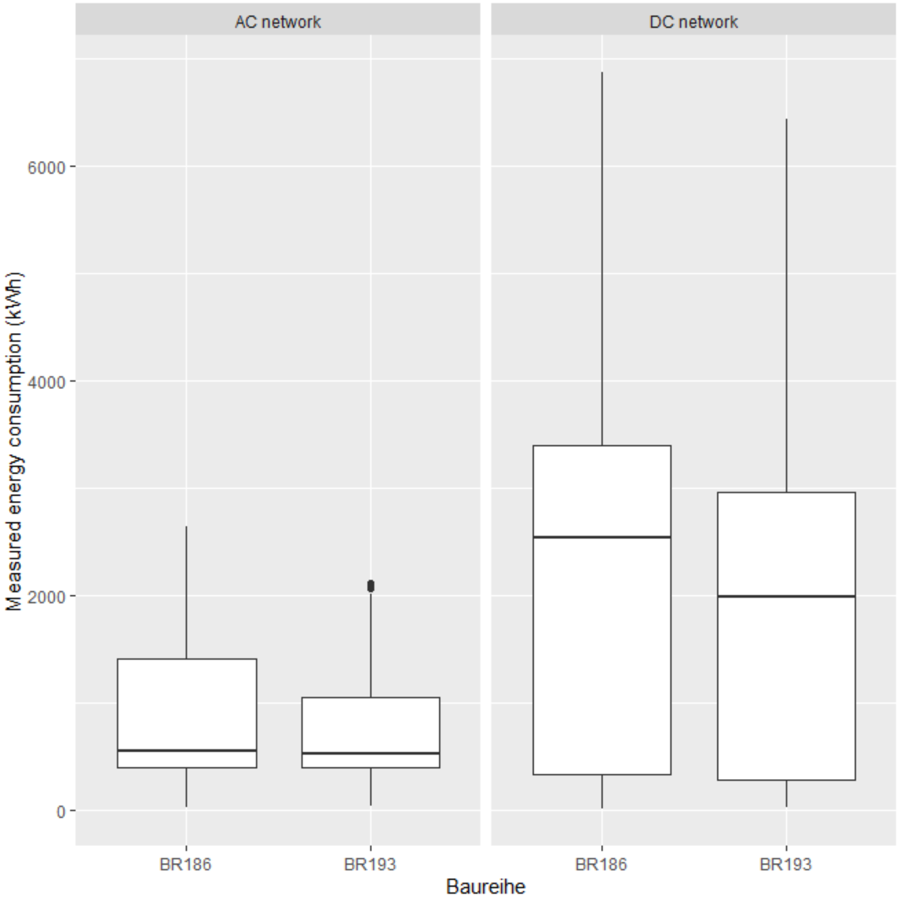


Figure 12: Boxplot distribution of measured energy consumption given types of locomotives and networks

In conclusion, we can see some trends by visual examination which indicate an increase in energy consumption when distance increases, with lower energy consumptions for transports executed on the AC network and with locomotives of type BR193. Interaction effects can have influence on the energy consumption and will therefore be included in the steps when building the model.

4.3.2. Analysis of current method and measurement data

As we have values for both the estimation from the current method and the measurements as sent by the locomotives, we can get a first indication about the accuracy of the estimations. By simply dividing the value of the measurement with the value of the estimations, we get the ratios as depicted in Table 1. It shows that the measurements in general are 87.6% of the estimations, indicating that the energy consumption is estimated at a higher value than the actual situation. Noteworthy is that the measurements for the grid “G-VL” is higher than the estimation. This goes against the results of the other grids. This can be explained by the characteristics of this grid and the way the value of the measured energy consumption is gathered. As the data is sent out by the locomotive every 5 minutes, and the grid section can be crossed within 5 minutes, it gives an inaccurate measurement as the allocated energy consumption consists of energy that has been consumed at a different grid. Therefore, the possibility is there that a high measurement compared to the estimation is caused by the characteristics on a different location.

Grid	Measurement in percentage of estimation
BR-A15	78.6%
BR-HVSPL	84.8%
CONV	89.8%
G-VL	102.6%
Overall	*87.6%*

Table 1: Measurement of energy consumption as a percentage of the estimated energy consumption

To dive deeper in the differences between locations, we consider a boxplot (Figure 13) for each of the abovementioned grids. We see for the grids on the AC network, being *BR-A15* and *BR-HVSPL* a lower measurement than the estimation in the majority of the cases. For the remaining grids, we observe that the estimations from the equation overestimate the energy consumption given the measurements when the distance and time travelled is long enough to be considered realistic, as we already discussed is not necessarily the case for the *G-VL* grid.

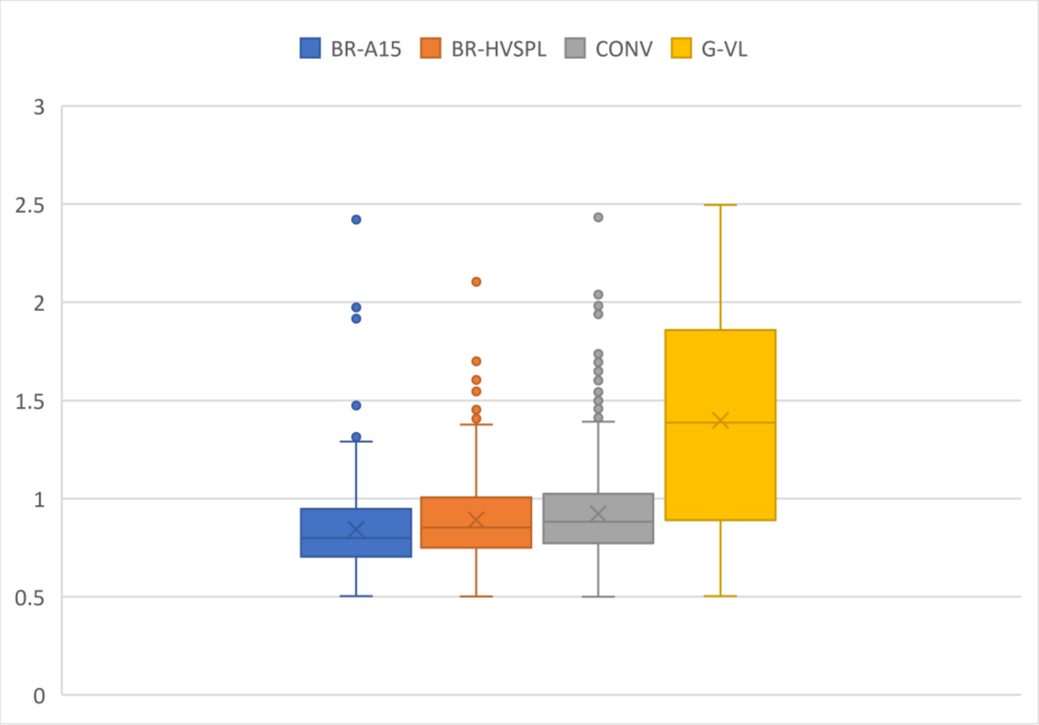


Figure 13: Boxplots of measurements expressed as percentage of estimations

5. Model training

The two methods we select for training a model are regression and random forest. Regression can result in a easy to understand final model which allows room for good interpretability while also possibly being accurate. Random forest lacks interpretability, but has the possibility for high accuracy. By comparing these two methods in addition to the model that is currently in place, we can determine whether the used estimation model is accurate or whether it can be improved.

Other techniques mentioned in section 3.4 were considered as well. However, we believe that they are either not suitable for this specific problem or lack accuracy in comparison to the chosen methods. Therefore, this research will only consider the aforementioned techniques.

5.1. Regression

First, the data is divided in a training and a test dataset, split 80% for training and 20% for testing.

To find which variables can influence the energy consumption, we first perform a linear regression of all variables included in the filtered dataset. The variables considered are:

- Net
- Weight (ton)
- Distance (km)
- Baureihe
- Months
- Part of Day
- Peak

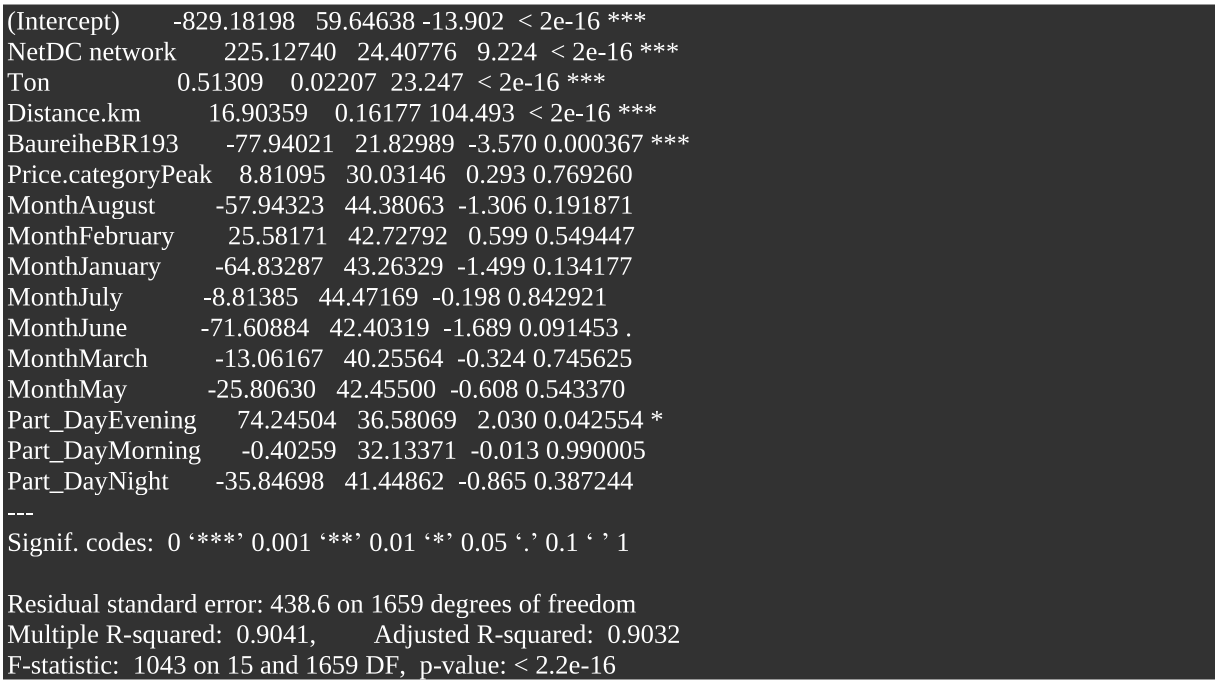
The grids have been dropped due to the characteristics being broadly similar on each network. Because of this similarity, a choice between network or grid has to be made so a model can be found.

The result of this regression results in the following:

```
Call:
lm(formula = Gem_Net_kWh ~ . - Grid, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2329.0  -229.1   -45.9   151.0  4004.8 

Coefficients: (2 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
```



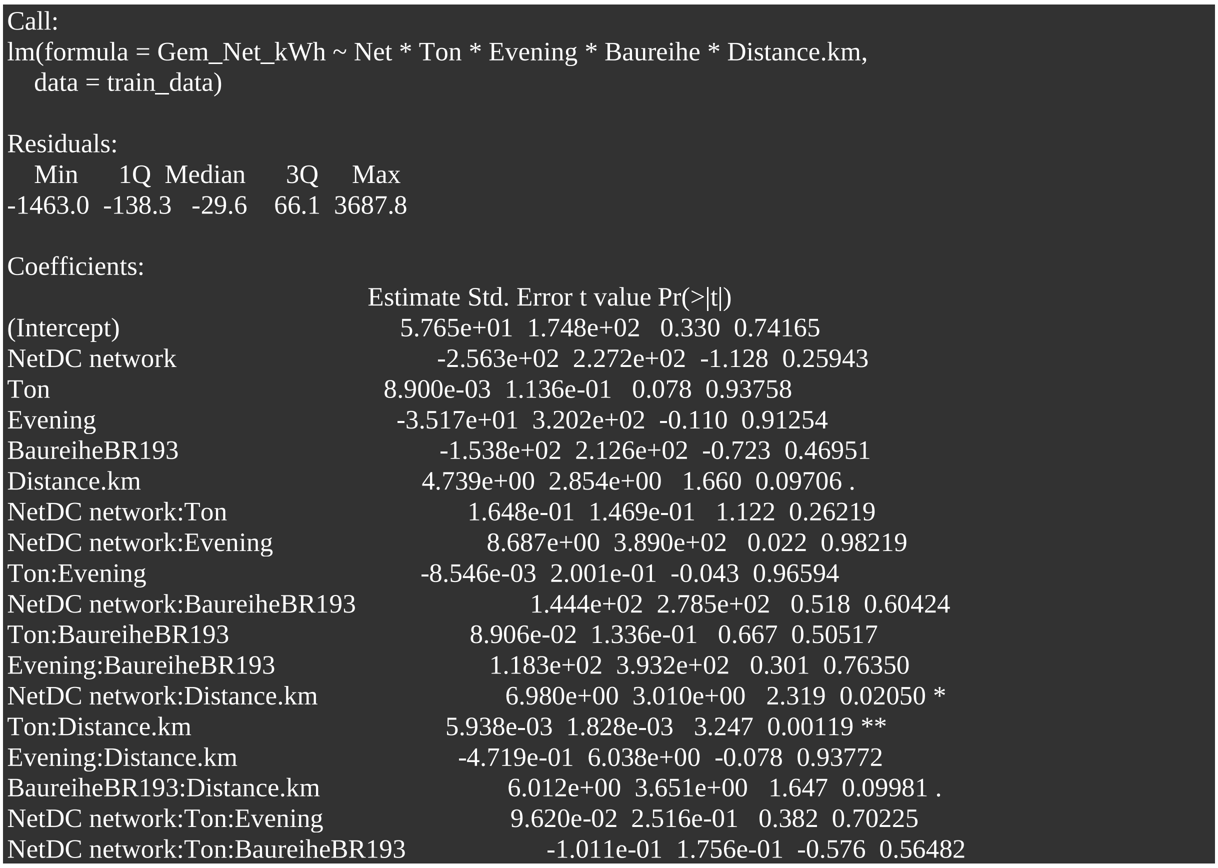
```
(Intercept)        -829.18198   59.64638 -13.902  < 2e-16 ***
NetDC network       225.12740   24.40776   9.224  < 2e-16 ***
Ton                   0.51309    0.02207  23.247  < 2e-16 ***
Distance.km          16.90359    0.16177 104.493  < 2e-16 ***
BaureiheBR193       -77.94021   21.82989  -3.570 0.000367 ***
Price.categoryPeak    8.81095   30.03146   0.293 0.769260
MonthAugust         -57.94323   44.38063  -1.306 0.191871
MonthFebruary        25.58171   42.72792   0.599 0.549447
MonthJanuary        -64.83287   43.26329  -1.499 0.134177
MonthJuly            -8.81385   44.47169  -0.198 0.842921
MonthJune           -71.60884   42.40319  -1.689 0.091453 .
MonthMarch          -13.06167   40.25564  -0.324 0.745625
MonthMay            -25.80630   42.45500  -0.608 0.543370
Part_DayEvening      74.24504   36.58069   2.030 0.042554 *
Part_DayMorning      -0.40259   32.13371  -0.013 0.990005
Part_DayNight       -35.84698   41.44862  -0.865 0.387244
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 438.6 on 1659 degrees of freedom
Multiple R-squared:  0.9041,	Adjusted R-squared:  0.9032 
F-statistic:  1043 on 15 and 1659 DF,  p-value: < 2.2e-16
```

Figure 14: Output of regression training without interaction variables

We set the cut-off percentage for significance at 5%. Therefore, we will drop the effects of peak and the month on the energy consumption as this shows to be insignificant. We add a binary variable to the dataset that shows if an entry took place in the evening (1) or not (0) to determine the effects this has.

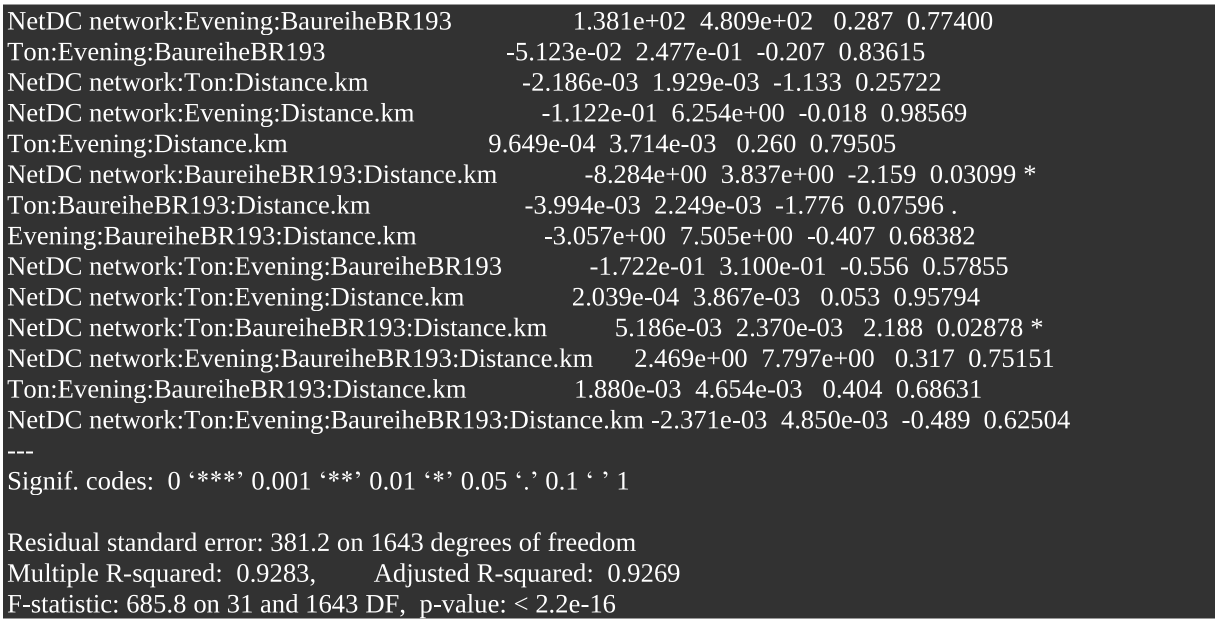
Re-running the model with interaction effects gives the following result:



```
Call:
lm(formula = Gem_Net_kWh ~ Net * Ton * Evening * Baureihe * Distance.km, 
    data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1463.0  -138.3   -29.6    66.1  3687.8 

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)  
(Intercept)                                    5.765e+01  1.748e+02   0.330  0.74165  
NetDC network                                 -2.563e+02  2.272e+02  -1.128  0.25943  
Ton                                            8.900e-03  1.136e-01   0.078  0.93758  
Evening                                       -3.517e+01  3.202e+02  -0.110  0.91254  
BaureiheBR193                                 -1.538e+02  2.126e+02  -0.723  0.46951  
Distance.km                                    4.739e+00  2.854e+00   1.660  0.09706 .
NetDC network:Ton                              1.648e-01  1.469e-01   1.122  0.26219  
NetDC network:Evening                          8.687e+00  3.890e+02   0.022  0.98219  
Ton:Evening                                   -8.546e-03  2.001e-01  -0.043  0.96594  
NetDC network:BaureiheBR193                    1.444e+02  2.785e+02   0.518  0.60424  
Ton:BaureiheBR193                              8.906e-02  1.336e-01   0.667  0.50517  
Evening:BaureiheBR193                          1.183e+02  3.932e+02   0.301  0.76350  
NetDC network:Distance.km                      6.980e+00  3.010e+00   2.319  0.02050 *
Ton:Distance.km                                5.938e-03  1.828e-03   3.247  0.00119 **
Evening:Distance.km                           -4.719e-01  6.038e+00  -0.078  0.93772  
BaureiheBR193:Distance.km                      6.012e+00  3.651e+00   1.647  0.09981 .
NetDC network:Ton:Evening                      9.620e-02  2.516e-01   0.382  0.70225  
NetDC network:Ton:BaureiheBR193               -1.011e-01  1.756e-01  -0.576  0.56482  
```



```
NetDC network:Evening:BaureiheBR193                1.381e+02  4.809e+02   0.287  0.77400
Ton:Evening:BaureiheBR193                         -5.123e-02  2.477e-01  -0.207  0.83615
NetDC network:Ton:Distance.km                     -2.186e-03  1.929e-03  -1.133  0.25722
NetDC network:Evening:Distance.km                 -1.122e-01  6.254e+00  -0.018  0.98569
Ton:Evening:Distance.km                            9.649e-04  3.714e-03   0.260  0.79505
NetDC network:BaureiheBR193:Distance.km           -8.284e+00  3.837e+00  -2.159  0.03099 *
Ton:BaureiheBR193:Distance.km                     -3.994e-03  2.249e-03  -1.776  0.07596 .
Evening:BaureiheBR193:Distance.km                 -3.057e+00  7.505e+00  -0.407  0.68382
NetDC network:Ton:Evening:BaureiheBR193           -1.722e-01  3.100e-01  -0.556  0.57855
NetDC network:Ton:Evening:Distance.km              2.039e-04  3.867e-03   0.053  0.95794
NetDC network:Ton:BaureiheBR193:Distance.km        5.186e-03  2.370e-03   2.188  0.02878 *
NetDC network:Evening:BaureiheBR193:Distance.km    2.469e+00  7.797e+00   0.317  0.75151
Ton:Evening:BaureiheBR193:Distance.km              1.880e-03  4.654e-03   0.404  0.68631
NetDC network:Ton:Evening:BaureiheBR193:Distance.km -2.371e-03  4.850e-03  -0.489  0.62504
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 381.2 on 1643 degrees of freedom
Multiple R-squared:  0.9283,	Adjusted R-squared:  0.9269
F-statistic: 685.8 on 31 and 1643 DF,  p-value: < 2.2e-16
```

Figure 15: Output of regression training without selected variables, including interaction variables

Again taking the significance of 5% gives the following (interaction) variables.

- Network * Distance
- Weight * Distance
- Network * Baureihe * Distance
- Network * Weight * Baureihe * Distance
- Network
- Distance
- Weight
- Baureihe

These are the variables we consider for the final regression model.

Coefficients:

Intercept	5.057e+01
NetDC network	-2.212e+02
Ton	1.584e-03
Distance.km	4.517e+00
BaureiheBR193	-1.370e+02
NetDC network:Distance.km	6.914e+00
Ton:Distance.km	6.350e-03
NetDC network: BaureiheBR193	1.980e+02
Distance.km:BaureiheBR193	5.795e+00
NetDC network:Ton	1.646e-01
Ton:BaureiheBR193	8.661e-02
NetDC network:Distance.km:BaureiheBR193	-8.237e+00
NetDC network:Ton:Distance.km	-2.125e-03
NetDC network:Ton:BaureiheBR193	-1.483e-01
Ton:Distance.km:BaureiheBR193	-3.914e-03
NetDC network:Ton:Distance.km:BaureiheBR193	4.897e-03

Table 2: Output of regression training model with coefficients

The resulting formula looks as follows:

$$Energy\ consumption = 50.57 - 221.2x_1 + 0.001584x_2 + 4.517x_3 - 137x_4$$

$$+6.914x_1x_3 + 0.00635x_2x_3 + 198x_1x_4 + 5.795x_3x_4 + 0.01646x_1x_2$$

$$+0.08661x_2x_4 - 8.237x_1x_3x_4 - 0.002125x_1x_2x_3 - 0.1483x_1x_2x_4$$

$$-0.003914x_2x_3x_4 + 0.004897x_1x_2x_3x_4$$

Where:

$x_1 = 1, if\ DC\ network, 0\ otherwise,$

$x_2 = Weight\ (ton)$

$x_3 = Distance\ (km)$

$x_4 = 1\ if\ Baureihe\ BR193, 0\ otherwise$

5.2. Random Forest

The second method we use to get a model for the prediction of the energy consumption is the random forest. We discussed what random forests are in the discussion of modelling techniques in section 3.4.

The resulting model cannot be visualized due to its complex nature. The model therefore remains a black box that provides a result when given a set of data. In comparison to the regression model, no additional data processing has taken place in any step of the model training as all variables are considered. The output of the model after finalizing the training is depicted below.

Random Forest

```
            Type of random forest: regression
                 Number of trees: 500
No. of variables tried at each split: 2
        Mean of squared residuals: 169058.5
                % Var explained: 91.49
```

A benefit of the random forest model is that the variable importance can be easily obtained. The variable importance graph shows which variables have the most impact on the final prediction within the decision trees used in the random forest model. It shows that the distance is by far the most important variable in this model, followed by Grid and Net. What is interesting is the relative low importance of the “Baureihe” variable, the type of locomotive.

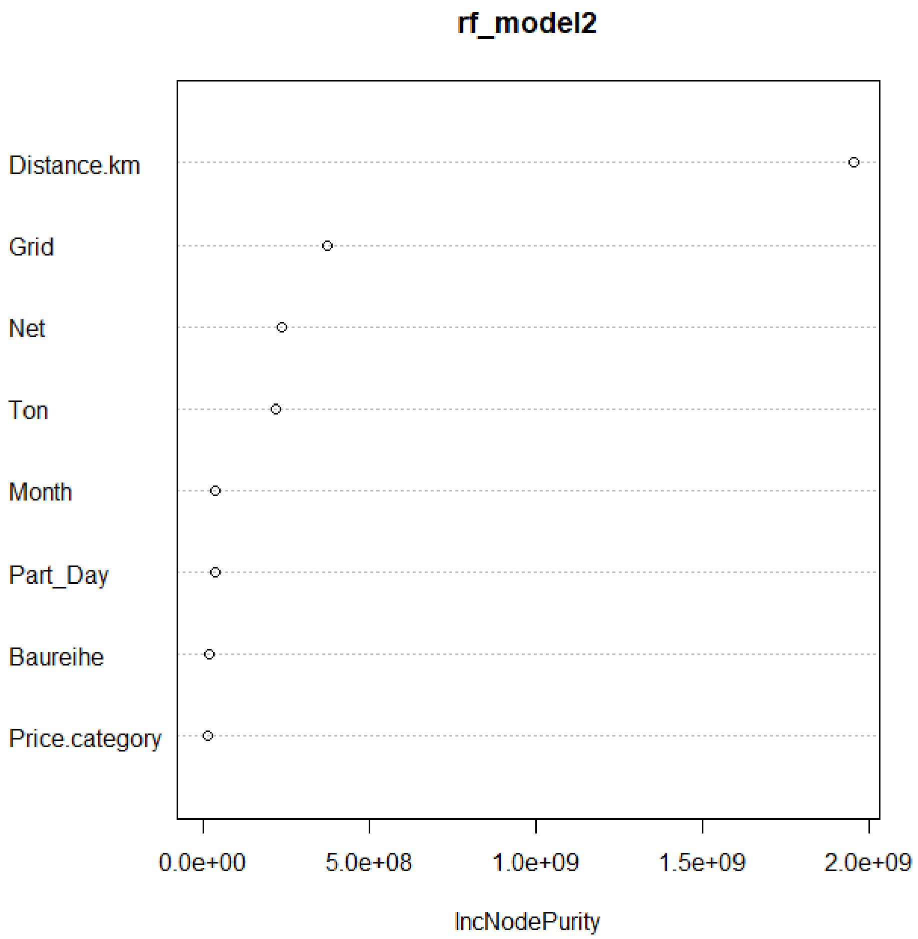


Figure 16: Variable importance of trained model using random forest

5.3. Model comparison

The goal of the research as stated in the beginning is to find the factors that influence the energy consumption of transports and to determine the accuracy of the prediction formula that is used to determine it. The first goal has been researched in the past sections. In this section, we will compare the different models to each other. First, we compare the models on a test dataset taken from the available data before training of the models. This data is used to predict the energy consumption on both models. For the real world prediction model, the estimation is known. The (root) mean square error can be calculated for each model, which are shown in Table 33.

Model	*MSE*	*RMSE*
Regression	131657.6	362.8465
Random Forest	140304.1	374.5719
Current Method	286439.3	535.2003

Table 3: Comparison of performance of three models on test dataset

Besides the total value of the (R)MSE, we can also explore the spread of prediction errors. These are depicted in Figure 17. The regression model has a limited spread of prediction errors, with the smallest box of the three models. The current method has a larger spread. Furthermore, the spread of the current method tends to be an overestimation in most cases, which can be seen by the largest part of the box being above the "0" line. Random Forest has a prediction error spread slightly larger than the regression model, but the predictions are more accurate than the current method.

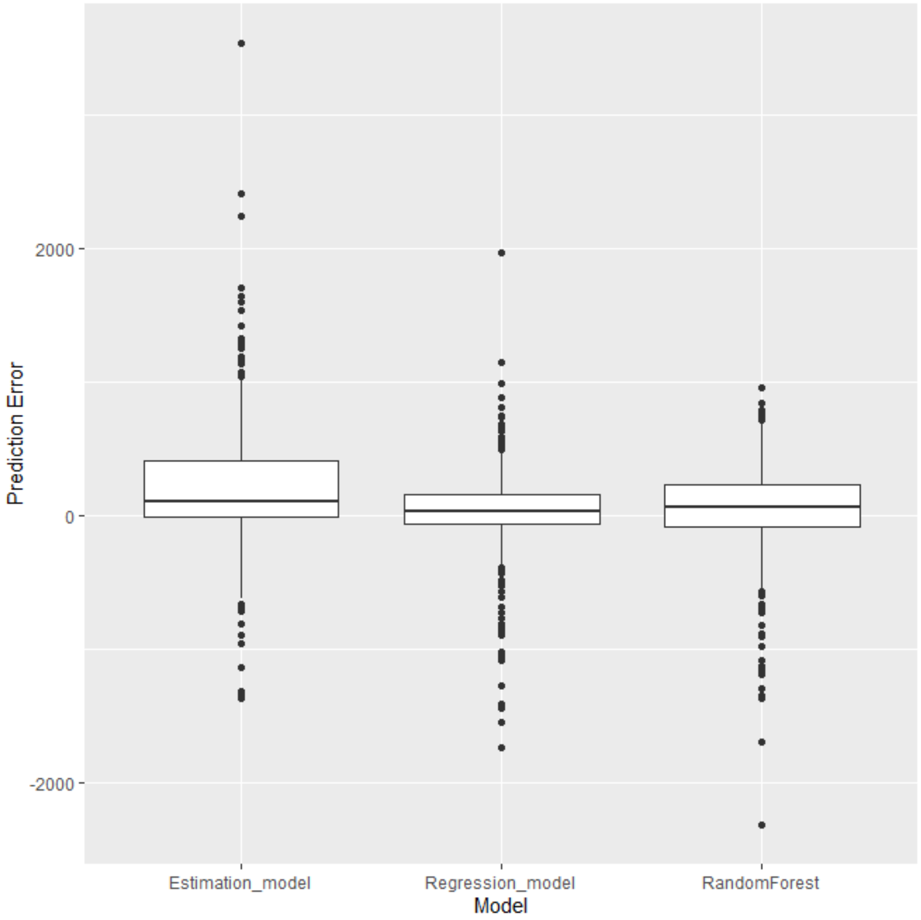


Figure 17: Spread of prediction errors for current method, regression model, and random forest

For further examination of the performance of the models, the effects of a changing variable are plotted for each model. We compare for three different grids the effect of increasing the weight and the distance for both types of locomotives taken into account in this study on the value of the prediction of the energy consumption. Additionally, the estimation of the energy consumption for a given combination is plotted. When increasing the weight, the distance has been set at 100 km. When increasing the distance, the weight has been set at 1500 tons. The graphs can be found in Appendix A: Model comparison.

It shows that the models obtained by training put out a higher predicted energy consumption at lower weights compared to the estimation model, while at higher weights the outcomes are lower. This could indicate that the estimation model overestimates the energy needed for the transport of heavy freights. This trend shows in comparison to both the random forest and the regression models in all combinations of grid and locomotive type.

When comparing the outcomes over distance, we see a similar trend where the current method predicts lower consumptions than the random forest model at smaller distances and similar consumptions as the regression model, but predicting higher numbers when the distance increases.

Based on the previous graphs, it would suggest that the estimation model overestimates the energy consumption. This could be a result of circumstance with the values chosen for testing in the previous comparisons. To give more certainty about the outcome of the models, we compare the performance of the regression model and the estimation model when both weight and distance increase. We do this on a 3 by 3 scale with a low/medium/high value for both weight and distance to limit the number of combinations. The regression model is chosen due to the better performance as per Table 3, as well as the better interpretability of the random forest. The values chosen for this test are depicted in Table 44.

Distance		**Weight**	
Low	10	Low	200
Medium	50	Medium	1200
High	150	High	2500

Table 4: Values of variables for evaluation of regression model

The results have been plotted and can be found in Appendix B: Test case for regression model. The outcome of this test show a few interesting things. First, the regression model allows for a negative prediction of energy consumption when both weight and distance are low on the direct current network. Furthermore, the depiction of higher predictions of the estimation models does not reoccur when the weight is low. In these cases, the regression model gives higher predictions as output. The same can be seen in cases where the distance is low, regardless of network. When neither distance or weight are in the “low” category, the prediction of the estimation model is similar or higher than the regression model.

In conclusion, the regression model performs the best on the test data of the three models that were compared. The random forest performed slightly worse, while the estimation model scored even worse. Upon further investigation, the regression model can give impossible outputs in certain scenarios such as negative energy consumption. Certain trends can be discovered when comparing predictions, such as high predictions of the estimation model when certain factors are increased. The outcome of the estimation model on the test data could be a result of overestimation in cases where the weight and distance of the transport are on the higher end, whereas the possible underestimation of the regression model in the cases where weight or distance is on the lower end is less severe.

6. Conclusion and Discussion

6.1. Conclusion

In the past sections, we have found that there are four main characteristics that influence the energy consumption, being the distance of the transport, the weight of the transport, the type of network where the train travels during the transport, and the type of locomotive that pulls the transport. Which variables are the most important for a model can differ slightly based on the method. Other potential factors present in the available data did not show any significant effect on the prediction of the energy consumption.

We have found a regression model that improves on the theory-based estimation model that is currently implemented for estimating the energy consumption. Given the comparison between the estimation and the measured consumption in chapter 4 and the comparison between the different models in chapter 5, we conclude that the estimation model generally overestimates the energy consumption. As the regression model proved to be more accurate over a randomly selected subset of the available data, the predictions from this model are generally lower than the estimation model.

The goal of this research for LTE Netherlands was to determine whether approving the measurements of locomotive meters for the payment of the energy consumption would be beneficial for the company. The resulting regression model in this research models these measurements and have shown to be lower than the estimations of the currently used model. Therefore, using the actual measurements would be beneficial as the used energy consumption is lower than the amount they are billed for. Our recommendation is to get the meters approved and used for the payment of the energy consumption.

6.2. Discussion

Although we have found that the regression model performs the best of all considered models, there is still room for improvement. For the scope of the research, this model suffices. However, it is not flawless. We already showed that for certain cases the prediction output could be negative. This is a very unlikely situation. A possible explanation for this is too little similar cases in the training data, resulting in an underfit of this area in the model. A different method of data selection could remedy this situation, as well as increasing the amount of data used for training. Additionally, a different model that performs better for these cases could be made.

The research has focused solely on the data that was available for finding the model. It got a little insight in which factors influence the energy consumption. However, it did not go deeper into the data to discover additional variables. Interesting variables such as accelerations and braking into standstills were not considered due to necessary data missing or not easily obtainable. Geographical characteristics such as inclines were also not taken into account to keep the focus on the scope of the research. Additional research into these areas could be interesting in order to get a fuller picture of what influences the energy needed to perform actions or overcome situations.

Within this research, we have kept the type of locomotive as a factor of its own. A locomotive is a complicated machine with its own characteristics. We have shown that the type of locomotive can influence the energy consumption. This does not delve deeper into which parts of a locomotive increase the energy needed or which part decrease it. Further research into this area could result into insights in the energy consumption of specific locomotives, which freight carriers can use to determine what locomotives to purchase, if energy consumption is a main trait they are interested in.

References

Bella, A., Ferri, C., Hernández-Orallo, J., & Ramirez-Quintana, M. J. (2010). Calibration of machine learning models. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* (pp. 128-146). IGI Global. http://dmip.webs.upv.es/papers/BFHRHandbook2010.pdf

CBS. (2022). *Prijs van energie 86 procent hoger*. Retrieved March 18th 2022 from https://www.cbs.nl/nl-nl/nieuws/2022/07/prijs-van-energie-86-procent-hoger

Heerkens, H., & Van Winden, A. (2017). *Solving Managerial Problems Systematically*. Noordhoff.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Nantasenamat, C. (2020). *How to Build a Machine Learning Model. A Visual Guide to Learning Data Science*. Retrieved June 6th from https://towardsdatascience.com/how-to-build-a-machine-learning-model-439ab8fb3fb1

Nespoulous, J., Soize, C., Funfschilling, C., & Perrin, G. (2021). Optimisation of train speed to limit energy consumption. *Vehicle System Dynamics*, 1-18.

Su, S., Tang, T., & Wang, Y. (2016). Evaluation of strategies to reducing traction energy consumption of metro systems using an optimal train control simulation model. *Energies*, *9*.

Wang, J., & Rakha, H. A. (2017). Electric train energy consumption modeling. *Applied energy*, *193*, 346-355.

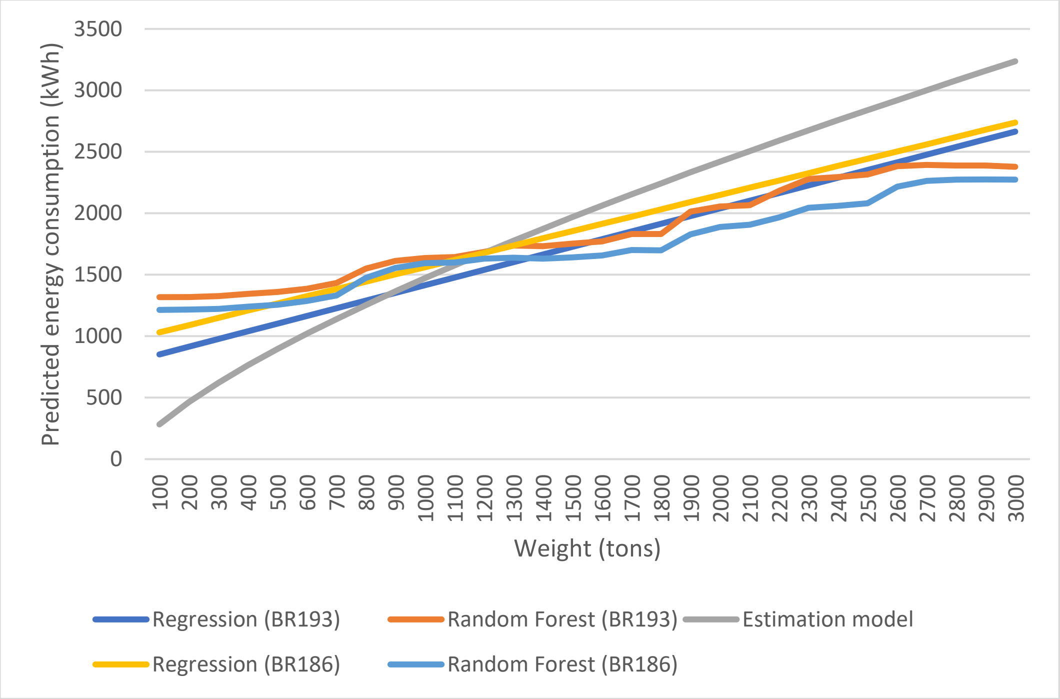
What is Machine Learning? A Definition. (2020). Retrieved June 6th from https://www.expert.ai/blog/machine-learning-definition/

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, *5*, 1205-1224. https://www.jmlr.org/papers/volume5/yu04a/yu04a.pdf

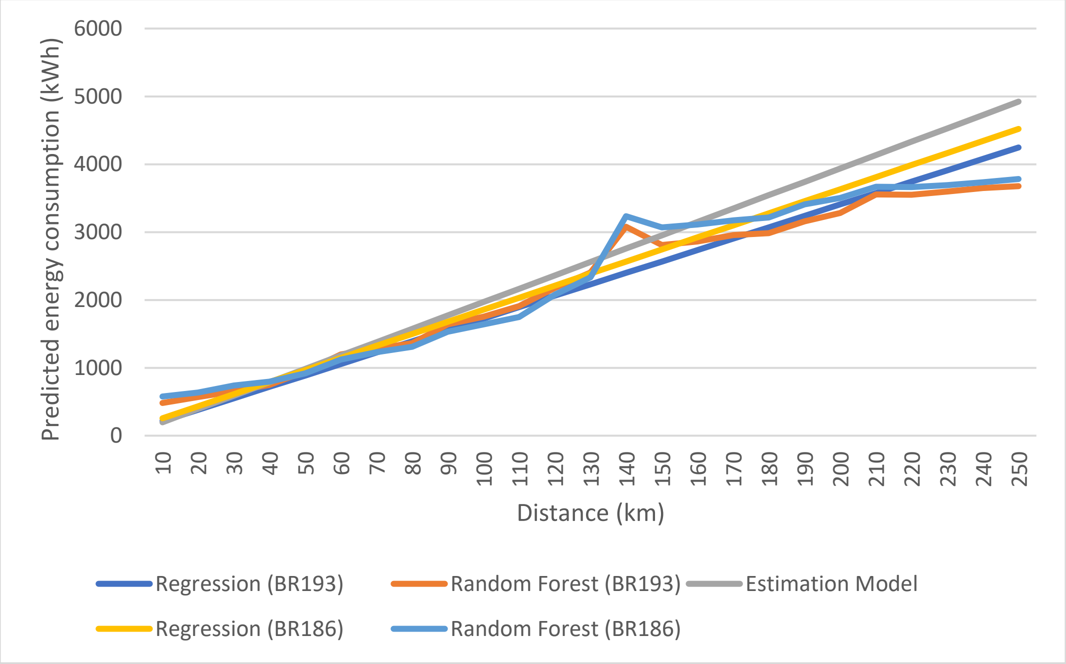
Yun, B., Baohua, M., Fangming, Z., Yong, D., & Chengbing, D. (2009). Energy-efficient driving strategy for freight trains based on power consumption analysis. *ournal of transportation systems engineering and information technology*, *9*(3), 43-50.

Appendix

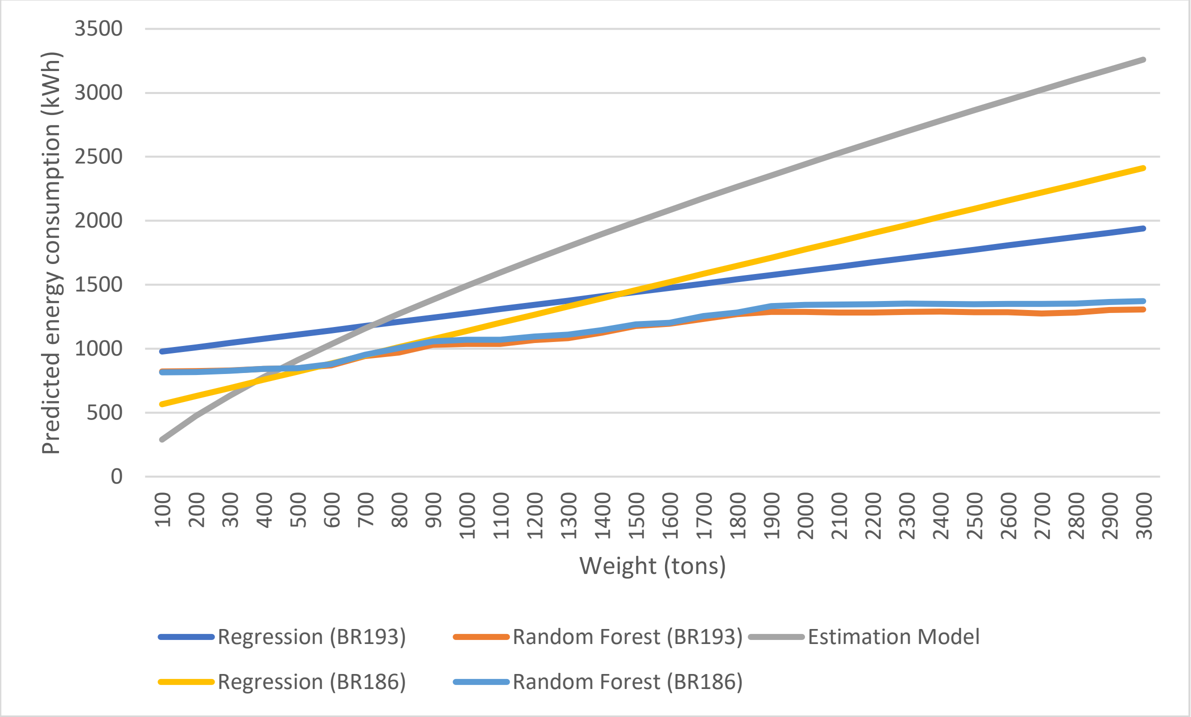
Appendix A: Model comparison



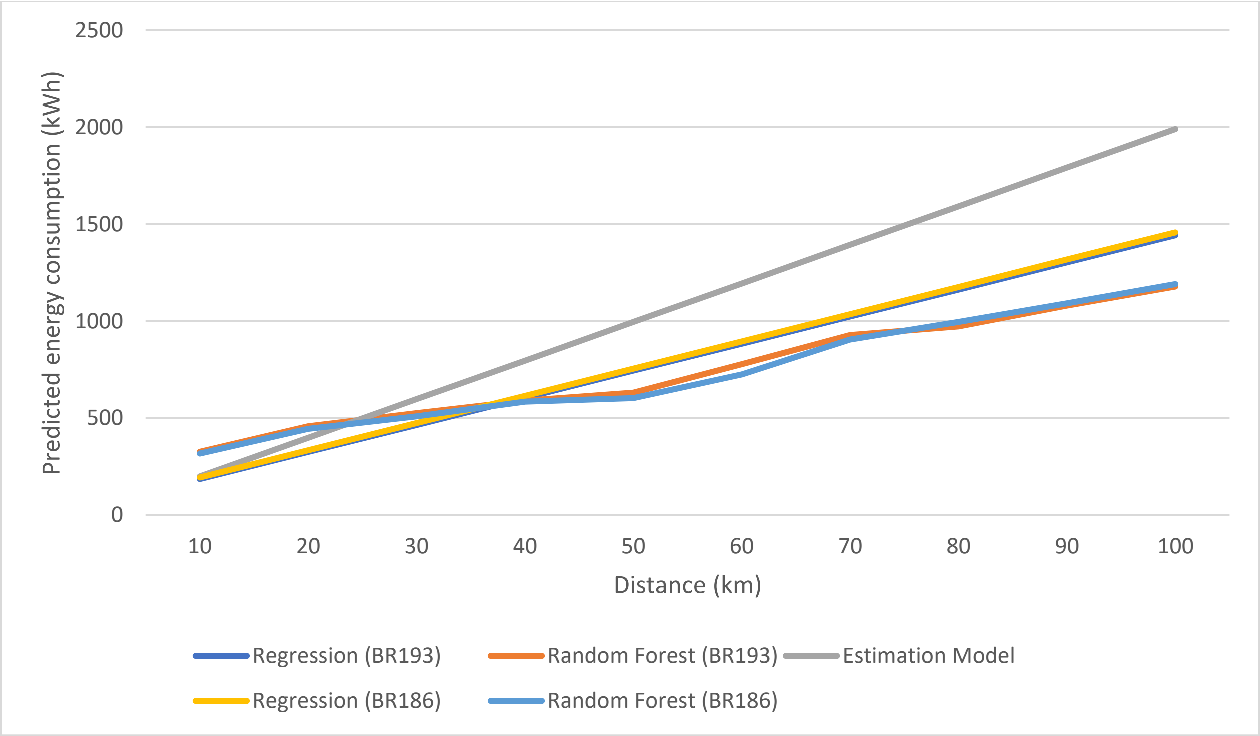
Appendix A- 1: Prediction of energy consumption for different models with increasing weight on CONV grid



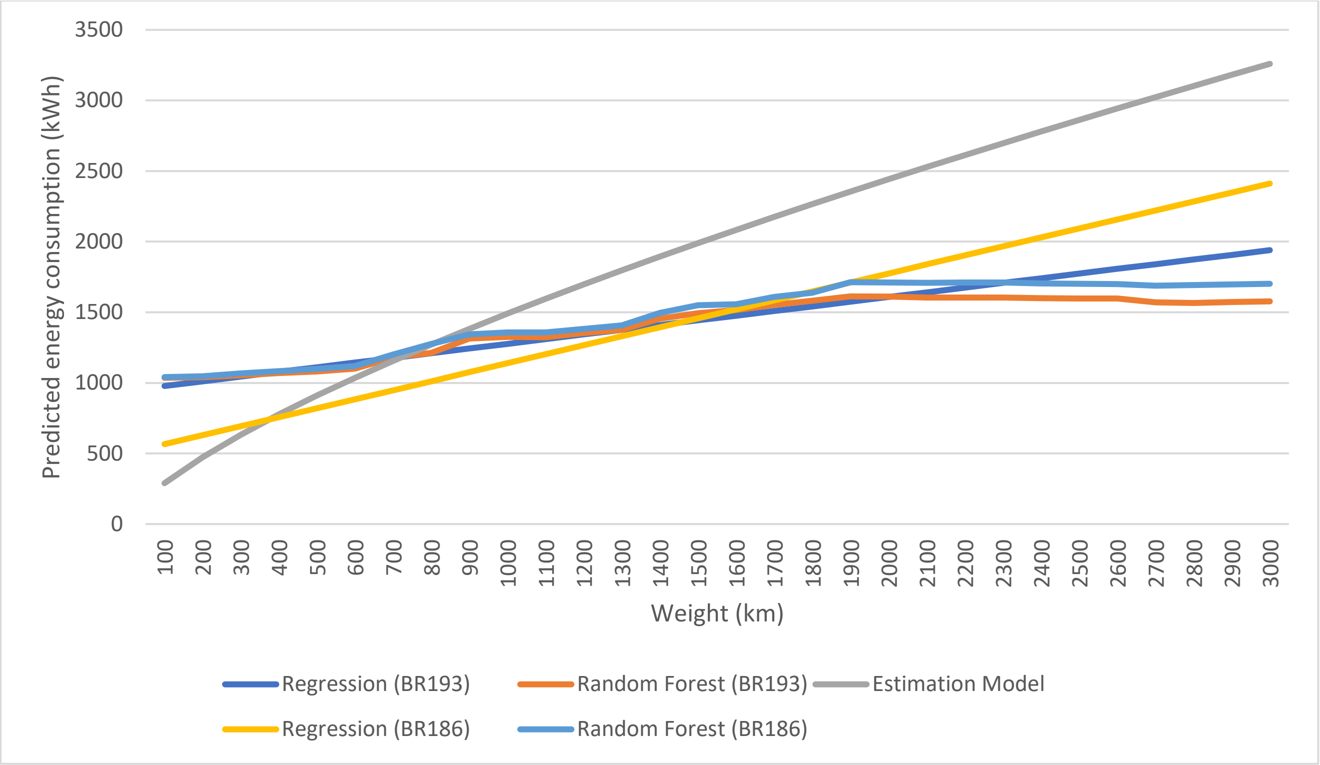
Appendix A- 2: Prediction of energy consumption for different models with increasing distance on CONV grid



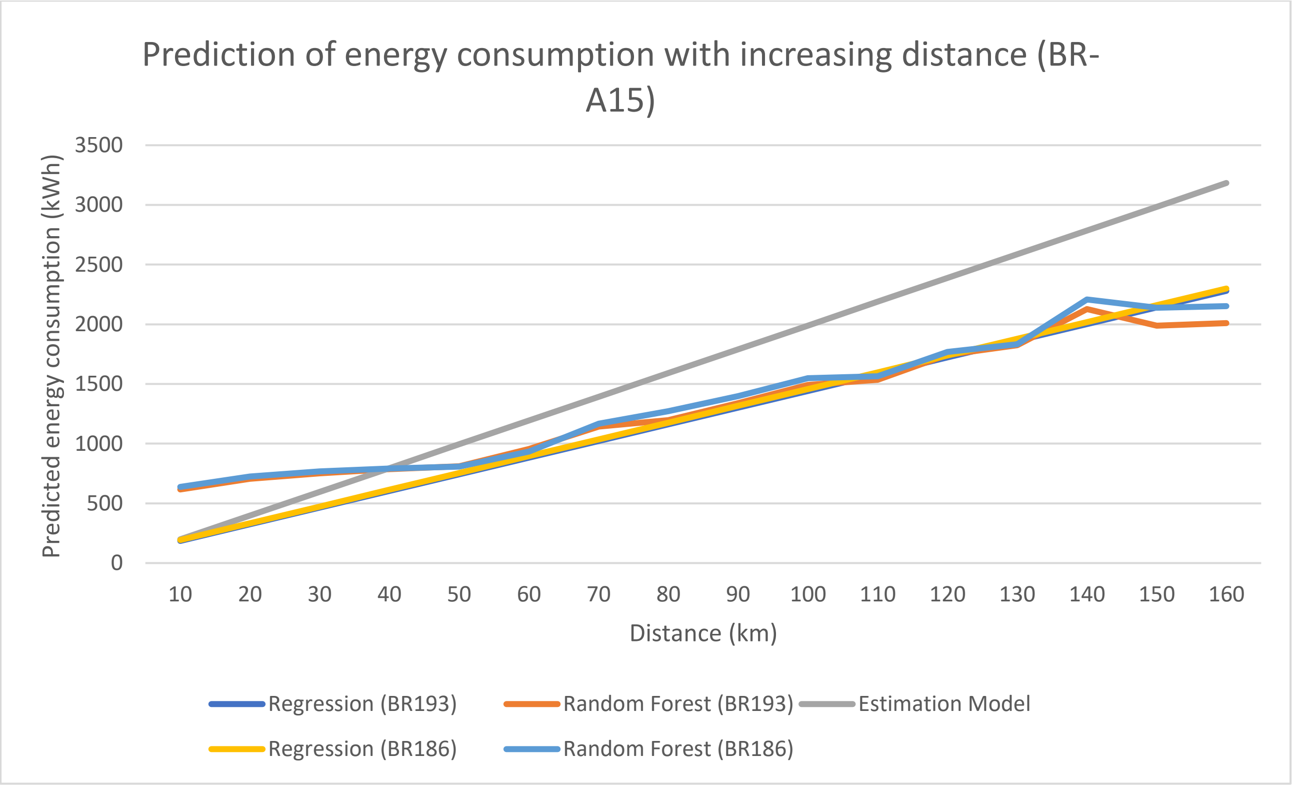
Appendix A- 3: Prediction of energy consumption for different models with increasing weight on BR-HVSPL grid



Appendix A- 4: Prediction of energy consumption for different models with incrasing distance on BR-HVSPL grid

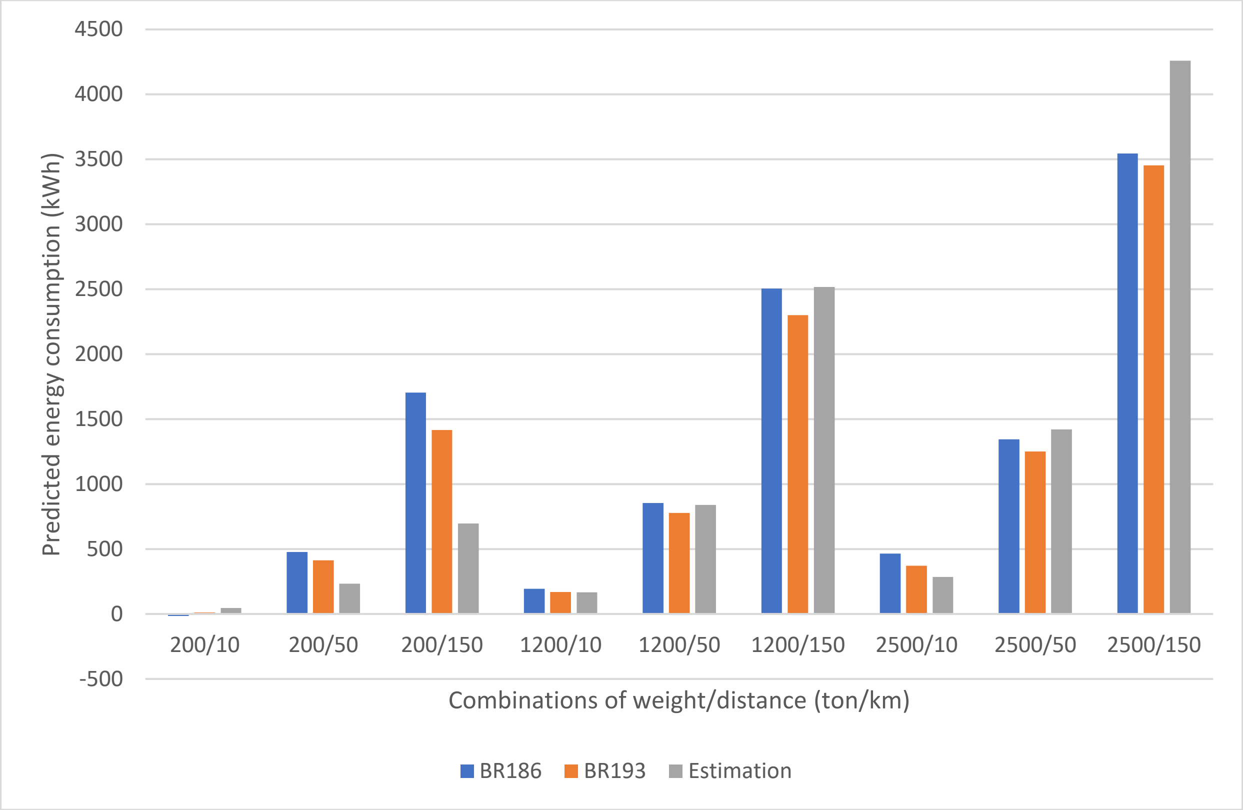


Appendix A- 5: Prediction of energy consumption for different models with increasing weight on BR-A15 grid

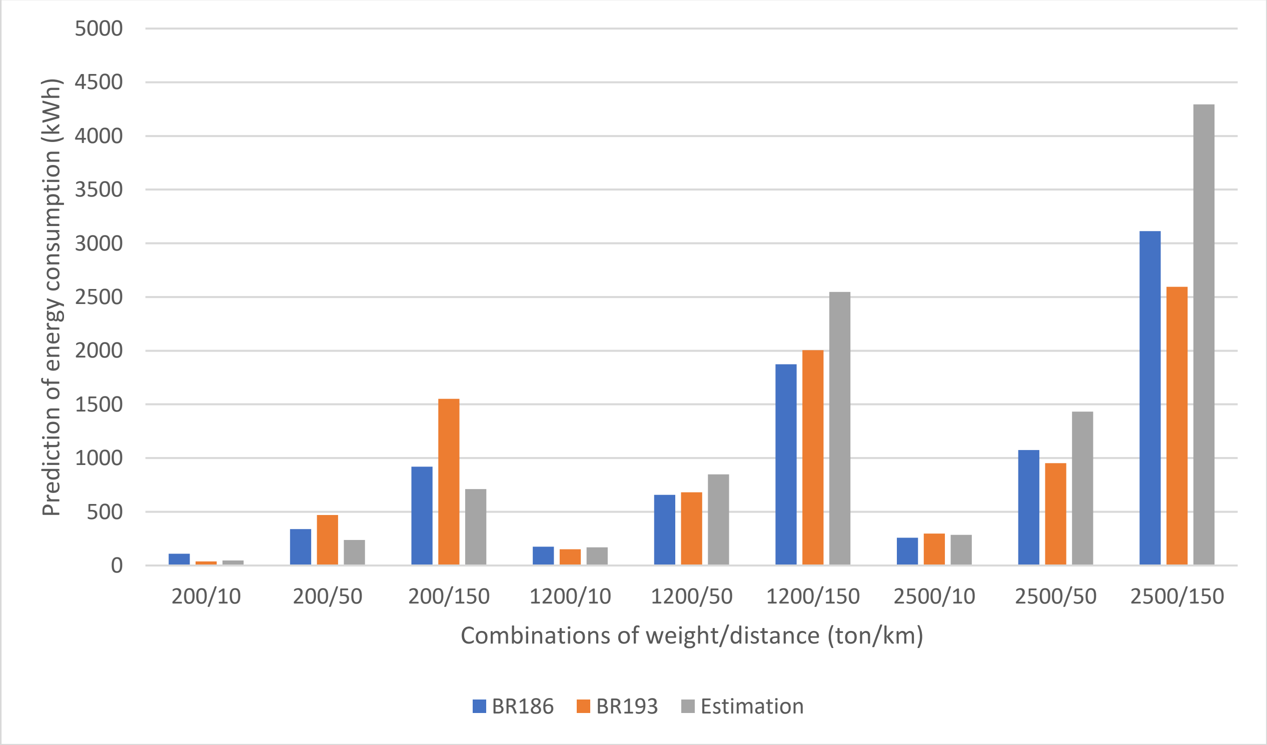


Appendix A- 6: Prediction of energy consumption for different models with increasing distance on BR-A15 grid

Appendix B: Test case for regression model



Appendix B- 1: Results of test case for regression model on DC network



Appendix B- 2: Results of test case for the regression model on AC network