University of Twente

Whether wetter weather is better

Determining the influence of short-term effects on the skid resistance

David van den Berg
Industrial Engineering and Management

This report is a summary of my research about influences on the measurements of skid resistance.

Q-Consult Progress Partners
Koeweistraat 1
4181 CD Waardenburg
Tel. (085)0 16 04 58

Rijkswaterstaat
Griffioenlaan 2
3526 LA Utrecht
Tel. 0800 8002

University of Twente
Industrial Engineering and Management
Postbus 217
7500 AE Enschede
Tel. (053)4 89 91 11

Determining the influence of short-term effects on the skid resistance

D.A.B. van den Berg
S1727478
Industrial Engineering and Management
University of Twente

Supervisors

University of Twente
M. Koot
Supervisor

University of Twente
Dr. IR. M.R.K. Mes
Second Supervisor

Q-Consult Progress Partners
Jan Telman
Senior Consultant & Trainer

Rijkswaterstaat
Frank Bouman
Senior Advisor

Rijkswaterstaat
Thijs Bennis
Advisor

Preface

Dear reader,

Before you lie my Industrial Engineering and Management bachelor thesis ironically called “Whether wetter weather is better”. The content is about the research I performed for Rijkswaterstaat regarding their correction model for skid resistance. I performed the analysis with the help of Jan Telman from Q-Consult Progress Partners. Thanks to him, I was able to understand, and adapt my regression analysis faster. Altogether I enjoyed our collaboration and the experience of working in professional business and working with an experienced consultant.

I also would like to thank my supervisor at the University of Twente, Martijn Koot, who helped me with writing my report and the methodology of data preparation. Thanks to his feedback and that of Martijn Mes, I am now able to present to you this report. I especially want to thank them for the feedback on my writing style because I was struggling with this.

With kind regards,
David van den Berg

Management summary

Problem Definition

To determine the quality of skid resistance of the national roads, owned by Rijkswaterstaat, the Side-way force (SWF) method is used. This method is used before by multiple companies and performed for every national highway each year. The process uses water during the measurement to simulate the effect of a wet road surface. However, the measurements are influenced by various short-term effects, which leads to variation in the skid resistance, expressed in SWF values. Currently, Rijkswaterstaat uses a model that corrects the measured SWF values to the expected SWF. Here the expected SWF is defined as a value that should be close to the SWF measured under standard circumstances.

However, the current model does not look at the influence of rain as a short-term effect. If this could explain the impact of seasonal variation used as a sinusoid in the model, a more accurate model could be formulated. Furthermore, Rijskwaterstaat wants to know whether the drought restriction is sufficient or could be altered. Therefore, the core problem is defined as follows:

There are unknown influences during and in advance of measuring, which negatively affect the accuracy of the correction model and the requirements to perform measurements.

Method

To solve this problem, we use measured data of previous measurements to investigate the influence of the variables. The goal of this analysis is to *formulate a model that corrects the measured SWF value as accurate as possible without making it too complex to function correctly.* We conclude that rain has a significant influence on the measured SWF method and was correlated with the seasonal variation. However, the influence of rain is too low to outweigh the added complexity it has on the model. A relation between drought and seasonal variation is determined. We developed new models to correct the measured SWF value and evaluated their accuracy, complexity and multicollinearity. In some of these models, the use of rain as a dummy variable, expressing occurrence, is used. The models with the best overall performance were used to investigate the necessity of the drought restriction. Since all the measurements met this restriction, we could only investigate if the limitation should be shortened or if the minimum amount of rain should be changed.

Results and discussion

Our recommendation is to use the following model using water and day number:

$$SWF_c = SWF + 0.0058 * (T_w - 20) - 0.0154 \sin(\frac{2\pi}{365}(x + 4))$$

SWF_c = Corrected SWF SWF = Measured SWF T_w = Temperature water x = Day number

This model is chosen since it has one of the highest accuracies and low in complexity and multicollinearity. The original model uses two temperatures and has the highest accuracy. Still, we concluded that the use of only one heat is enough and necessary. For the restriction of drought, we did not find any reason to change it. In a few cases, a small difference in corrected SWF values, between groups that met a shorter drought restriction, is determined. However, the increase in reliability does not outweigh the decrease in periods that measuring is allowed. To determine if the restriction can be shortened, we first need to assess the influence of rain on a short-term period, measurements should be performed where the requirement of drought is **not** met. Therefore, no good comparison and analysis with drought can be made in this report.

Table of content

1. Problem Statement

In this report, a research is done on the influences of multiple factors on the skid resistance of measured roads in the Netherlands. In this part, the problems of measuring the skid resistance are introduced. Based on these problems, a research approach has been made to help solve these problems.

1.1. Problem Introduction

Skid resistance is one of the quality indicators for Rijkswaterstaat (RWS) to measure the quality of the national highway network. Other indicators are raveling (surface damage due to loss of stones), track formation (due to the tires of heavy traffic), longitudinal flatness (bumps in the road) and cracking (collapse of the way due to, for example, subsidence of the subsoil).

These quality indicators are essential for RWS since they are the executive organization of the ministry of infrastructure and water management in the Netherland. The task of RWS is to work daily on securing and improving the safety, livability and accessibility in the Netherlands[1].

Skid resistance is found to be the most direct characteristics related to the safety of using the road. Skid resistance is measured as a coefficient of friction, indicating whether braking on the road can be done sufficiently. To measure this friction, it is important to use a system which does not obstruct the on-going traffic. For this reason, it is not practical and dangerous to use the braking distance of a car for each hectometer of road.

Figure 1-1 The sign on the right is placed when the road surface does not satisfy the legal requirements

Over the years, the road quality lowers continuously due to the traffic. The leading cause for this decrease is small stones, which polish by the passing tires of the traffic. Low skid resistance cannot be stopped or easily increased; in cases, the friction is below the legal set value, the surface needs to be replaced. In the meantime, the sign of Figure 1-1 is placed.

1.1.1. Measuring method Side way force

The measurements to determine the friction are performed by multiple companies, all using the same established method. Every year about 90,000 hectometer sections, which is about 9,000 km of road, is measured to determine the skid resistance and thus quality. This method requires a truck to drive with a speed of 80 km/h over the road. The friction is measured by pulling a specially prescribed unprofiled measuring tire along the road at a small angle (15 degrees). This is called the Side Way Force (SWF) Method. During the measurements, a small layer of water is added to the road. This layer of water is used to find the value of the skid resistance during rainy/stormy weather (when the friction is lowest).

Figure 1-2 Truck performing the SWF method

The skid resistance (expressed as an SWF-value) is a coefficient between 0.50 and 0.90 on national roads. The limit for the friction of a road is based on the accident risk. Currently, the limit of the friction is set on 0.51. A sharp increase in the number of accidents is found to be in hectometer sections with a coefficient of less than 0.51. [2]

1.1.2. Previous research

In the last years, research has been done to the influences of temperature on the skid resistance values found through the SWF method. This research concluded a considerable influence of *water-* and *road surface* temperature on the SWF[3, 4].

In a sequential research, the influence of seasonal factors has been measured. This concluded that the friction would be higher in March-June and lower in September-November. The cause for this seasonal effect is not yet known. It is expected that the surface is rougher after the winter because of the frost thaw cycles and spreading salt against the frost. The skid resistance lowers again in the summer because of polishing. At the same time, the temperature is influenced by the seasonal effect. In Figure 1-3[1], the relationship between the measured SWF and the day number it was measured is shown.

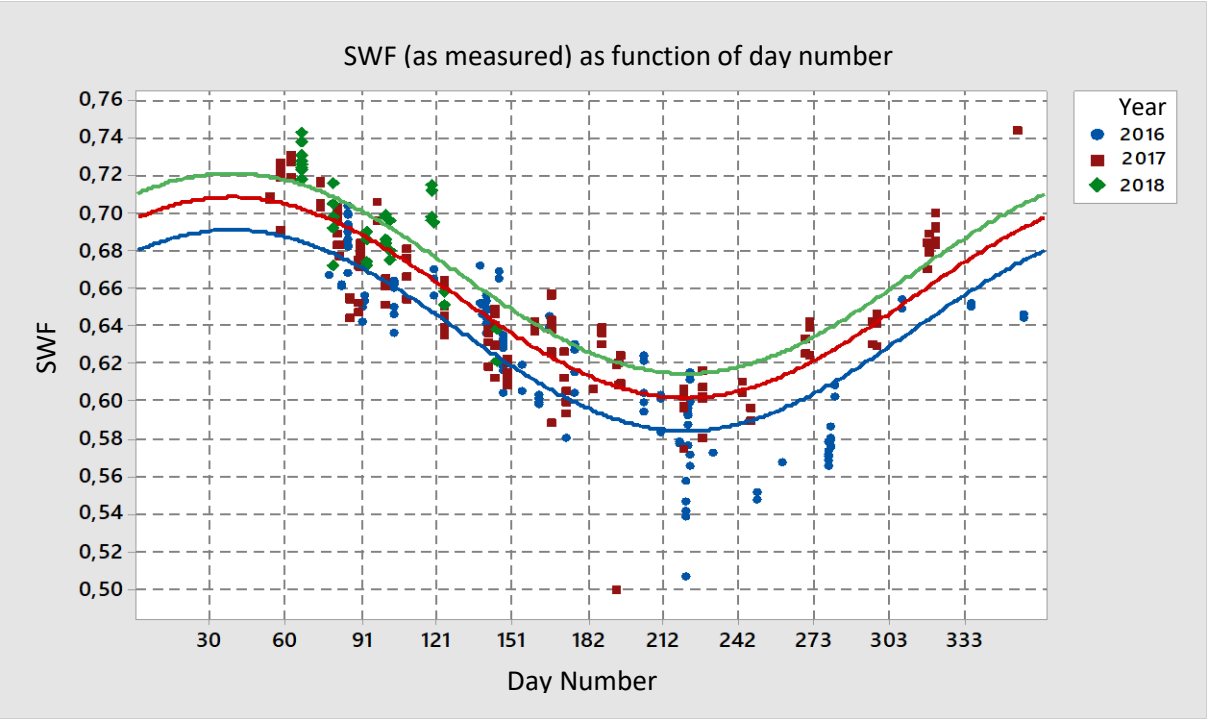

Figure 1-3 SWF measured values as function of days

In this scatterplot, the day number is presented at the X-axis, the found SWF (not corrected by the correction formula) is shown at the Y-axis. In the graph, three different sinusoids can be seen. The phase shift and amplitude of the sinusoid are calculated for three years. Each year has its baseline, and this is a stable value comparable to the expected SWF value during standard circumstances.

In the sinusoid, the influence of multiple factors is expressed if they are correlated with the day number. For example, the temperature is correlated with the date; therefore, if only a sinusoid is used to correct the SWF. This influence is part of the calculation. The sinusoid is based on the measurements over the year and therefore only a prediction for a specific day. After the correction of the temperature and sinusoid, there is still a remaining noise in the measurement data.

The goal of this research is to find and determine the remaining influences of weather conditions on the SWF measurements, which could reduce the noise of the difference between the corrected SWF value and the expected SWF value, by narrowing the standard deviation with the help of restrictions. Thus, to improve the accuracy of the model.

It is important to note that day number itself does not influence the SWF. Day number only predicts the deviation by seasonal influences; it is based on the correlation between date and measured SWF.

During the next year (2017) and the start of 2018, the same hectometer was measured. In the graph, the friction found with the SWF method (not corrected) as a function of the day in the year is shown for 2016 till 2018. These measurements of the same hectometer were measured to verify if the measuring instrument function properly. For this verification, it is assumed that the friction stays almost the same. Lower or higher measurements could be explained due to temperature or other influencing factors if the machines work properly. Therefore, the verification would identify a problem if there were weird fluctuations during these measurements.

The **small increase in SWF per year** should be noted since the theory states that the skid resistance should slowly decrease over time. However, a slight increase **may be** the **result of higher temperatures, less drought or other influences** that can differ each year.

1.1.3. Current formula

In the last research, a method was formulated to transform the measured SWF value to the expected SWF value, SWF corrected. To modify the data, the first input needed is the day number. Based on the day number, the influence of the seasonal factor can be excluded by subtracting the sinusoid. In the next step, the initial SWF value is corrected by the input of the water and road temperature. After this correction, the fixed measurements should be closer to the actual value. More information about the assumptions about adjusted, expected, and actual value can be found in Chapter 3.

$$SWF_c = SWF + 0{,}0035 \times (T_{water} - 20) + 0{,}0008 \times \left(T_{weg} - 20\right) - 0{,}0217 \times \sin\left(b(X + 15{,}2)\right)$$

SWF_c = Side way force corrected. The expected SWF at normal circumstances.

SWF = Side way force. The SWF found by the measurement system.

T_{water} = The temperature of the water used by the measurement system. This is 20°C at normal circumstances.

T_{weg} = The temperature of the road during the measuring. This is 20°C at normal circumstances.

X = Day number representing seasonal effect. This is 15 June, day 166, at normal conditions.

One of the goals in this project is to improve the accuracy of this formula. To reduce the number of outliers and increase the accuracy, additional input variables and their influence must be determined. The values are corrected to their expected values during normal circumstances. During normal conditions, the temperatures should all be 20 degrees, the seasonal influences should be the same as we would expect at day 166.

1.2. Problem Identification

In this chapter, the goal is to identify the core problem. For this identification, the method of Hans Heerkens is used as an inspiration[5]. In this report, the identification will be made in four steps. The first step is to create an inventory of the existing problems; in our case, this is the initial problem. After this, these problems will be made into a problem cluster to evaluate them in cause and effect. The third step is choosing the core problem, making it quantifiable.

1.2.1. Introduction

In this report, the importance of the model that RWS uses to evaluate has already been discussed. For RWS it is crucial to retrieve and correct the SWF values as close as possible to the actual values. For this correction there currently is already a model, this model is developed and based on the measured data of roads in the Netherlands. In principle, all hectometer sections are measured once a year. Since there are only three companies who perform these measurements, it is not possible to measure all road sections on one day. In these measurements, not only the date and time differs but also other variables which influence the measured SWF.

For a regular assessment of the roads, it is undesirable to have random circumstances influencing the friction on every location. To solve this, the correction formula is formulated. However, there is still a difference between the corrected SWF values of the same measurement place. Thus, the outcome of the correction still varies from the expected SWF values.

1.2.2. Lining up the problem

In the introduction, the original problem has already been mentioned. The use of an indicator for seasonal variation and lack of information about rain and drought lead to some incorrect corrections. This was the original problem and reason for the development of a correction model. Since it is impossible to recreate the same environment and circumstances for every measurement, a model is needed to convert the measured values to values at “normal” circumstances.

The current model is based on the difference in SWF and variables compared to the standard circumstances. The problem is that identifying the influence of a variable is hard since we cannot just change one variable to measure its impact on the SWF. Therefore, the influence of each variable is determined by the difference between normal circumstances and standard. The expected SWF is also found during this calculation. It is assumed that the expected SWF is nearly the same as the actual SWF (Chapter 3).

In some periods, the corrected SWF value is not close enough to the actual value. The next step is to identify why the accuracy of the current model is off. This can be evaluated in the amount of corrected values within acceptable margins of the actual value.

The first problem is the action problem. It is noticeable that if the model is used on some of the data where the actual value is known, the corrected value differs. Another reason for this research is that RWS wants to be sure that the model is correct (reliable corrections). There are no other models that can correct values from the SWF method. SWF is also used in Germany, but they have another type of road surface material.

1.2.3. Cause and Effect

The original problem and reason for the development of the current model was the significant variation in repeated measurements on the same road sections. With a correction model, these values are significantly improved to more comparable values. However, with newfound values it is noticeable that during certain timespans, 3 to 10 days, almost every corrected measurement differs from the actual value. A cause for this can be that the formulated model is based on previously found data. Therefore, the new data is done with a different circumstance which influences the measured SWF. A problem for determining the influences is that all the variables could be interdependent. Therefore, it is hard to find the exact impact of each variable, and there might be an additional influence of two variables on the formula.
Another problem that leads to inaccuracy in the results is the use of the sinusoid. The sinusoid is an indicator of the expected seasonal influences of a specific day. The sinusoid only predicts what the influence of these factors should be; it can lead to inaccuracy if the conditions differ from expected.
All these problems together form a (untraditional) problem cluster, see Figure 1-4. In this cluster, there are a lot of issues that influence each other. An additional action problem is that the model is based on measurements of random situations. The high correlations between temperatures leads to uncertainty for the best predictor. The current model uses water and road surface temperature, and these are correlated with the outside temperature. Therefore, RWS has already agreed to research this.

1.2.4. Choosing the core problem

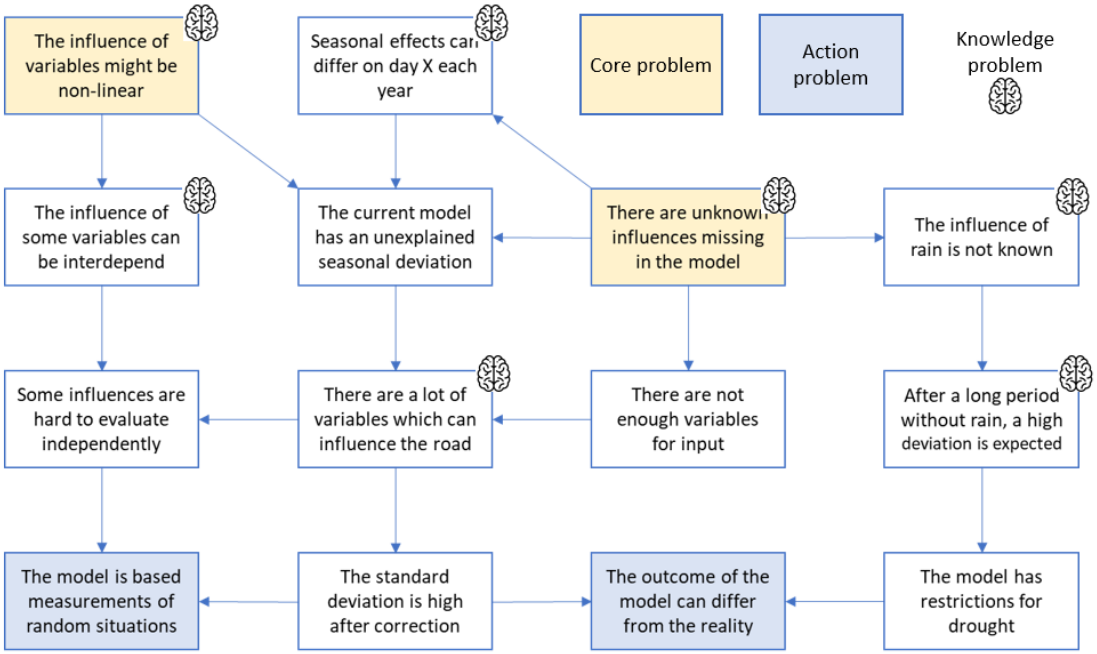

Figure 1-4 Identifying the possible problems.

Instead of the traditional problem cluster, we try to identify improvements that could have the most influence on the action problems. The first core problem could be that the model has not enough input variables for the correction. To solve this, we can determine the influence of more variables, adding those to the model and evaluating the restrictions that currently is used. The second core problem is that currently, there has only been done investigation to linear regression for the influences. Till now, only linear regression analysis is done because the calculations are based on the measurements of random situations.

Therefore, it is hard to investigate if the influences are also interdependent or perhaps be correlated in another way. Rijkswaterstaat recently has agreed to do research to the correlation that water temperature and road temperature have on each other. This research is necessary since the temperature of water and road surface both significantly are influenced by the temperature of the air. This is logical since the temperature of the air is the actual temperature outside and changes both these temperatures. This, however, makes it harder to determine the influences independently of each other. To measure this, they will perform measurements with cold water on a hotter road. This research can help with solving the second core problem since it more clearly shows the kind of relation between the variables and their SWF value.

For this report, the first core problem, the unknown influences missing in the model will be the primary focus. The reason for this is that it is still hard to evaluate the regression. This will be easier when the research for the second action problem is done. Another reason is the unknown influence of rain and drought. The reason for a restriction, which can hinder measuring, should be researched and only used if it is necessary.

The action problem and reason for this research are that the current model contains some flaws. The model can be optimized by detecting which variables are missing and determining what their influence on the friction is. The process of this optimization can be evaluated. For this report, the evaluation will be based on the confidence interval, residual noise and standard deviation.

An important note for this improvement is that the new model should not be too complex to use. This means that it should be easy to find the input variables needed for the correction. Then a trade-off can be made between the complexity and accuracy of the model.

1.3. Research Questions

For the problem approach, research must be done to gain more information. This helps with the possible methods and needed information to answer these questions. A goal of this report is to answer the research question; this question is based on the core problem. For the core problem it is stated that the influences of more variables need to be determined. These new variables help to formulate a new model that can correct the SWF value more accurately, this solves the initial action problem. An additional requirement is to keep the data needed for the model easy to measure, this keeps the model usable and not too complex. With all this information, a research question can be formulated to help make an improved model. The goal is to decrease the mean difference between the corrected values and the accurate values. This will be evaluated by the amount of corrected values that are within margin of the expected values and the standard deviation.

1.3.1. Research question

Which input variables and restrictions regarding weather variables should be used to form a model to correct the measured side way force as accurate as possible, without making it too complex to use it properly.

Based on the research question above, multiple sub-questions are formed. These sub-question help to answer the research question using the CRISP-DM methodology[6].

1.3.2. Sub questions

1. **How to determine the influence of each independent variable**
   - What are possible short-term influences on the measured SWF
   - How to specify the potential independent variables in their measurability
   - How to evaluate the correlation per variable

The answer to this question helps to choose the possible input variables for the model. There are a lot of options to measure for example, rain; this research contributes to pick the best. These specified variables can be used to determine their influence. After choosing the variables and finding their impact, the next step is to evaluate them.

2. **Which independent (input) variable should be chosen for the model to correct the SWF accurately?**
   - What makes a model too complex to work correctly?
   - How much work does it take to measure the data needed as input?
   - What is the influence of each variable on the outcome?
   - How to determine whether the complexity of the formula outweighs the significance on the outcome?

This answer helps to select the right variables from the options that are found in the first question. For the selection, research must be done about the evaluation of the variables. Based on this, an assessment can be made between the added complexity of using the variable and its influence on the outcome. After this step, the options can be rated and used to select the variables to include in a model. The next step is to choose the best model.

3. **How to choose the best model for correcting the SWF?**

- How to evaluate a model? (what aspects are essential)
- What is the advantage of each model?
- What is the disadvantage of each model?

The purpose of this question is to combine all the found information to formulate a new, improved model. In the first part of this question, it is essential to evaluate all the models. This has already been done in the third sub-question for accuracy, reliability and complexity. Still, this evaluation can include even more aspects. After evaluating the models, the important elements of a model for each stakeholder will be evaluated. For instance, the accuracy is more critical for RWS. Still, the complexity for the input variables is an essential aspect for the measuring companies. This may lead to multiple "best" models. In this part, their strengths and weaknesses will be explained.

4. **What can be done to make the model as reliable as possible?**
   - Is the current restriction sufficient?
   - What restriction can be added/changed to increase the reliability of the model?
   - What is the confidence interval of the model?
   - How to make a trade-off between reliability and practicality.

This question helps to choose the best variables based on the reliability they have. This starts with evaluating their reliability and then evaluating the reliability they would have in the possible models they can form. To make the model as accurate as possible, a small confidence interval would be suitable. Based on these findings, restrictions or other requirements can be added to the model, for example, measurements after two weeks with less than 1mm rain in total are not allowed.

1.4. Problem approach

After the correction of temperature and the seasonal effect, there is still a significant uncertainty in the SWF-data. This is visible by the scatters of measurements found around the sinusoid in see Figure 1-3.

It is known that the skid resistance is influenced by precipitation. There are also other possible weather factors which influence the friction. Drought is one of these factors. After a long period of drought, the skid resistance is lower when it rains again for the first time. This, therefore, also applies to the measurement, where water is also sprayed for the measuring tape. The pollution that is not regularly washed away may play a role in this. For this reason, RWS has imposed a restriction on the measurement companies: measurements may not be taken after a more extended period without precipitation (approximately two weeks).

It is desired to investigate the effect of precipitation on the friction value and SWF method to test the limits of days of drought. This can give more insight into the necessity of the drought restriction for the measuring companies. The focus will, however, be on all the variables. The addition of a variable also changes the influence of the current variable that is used. The reason for this is that variables can influence each other, or the addition can lead to overfitting.

1.4.1. Stakeholders

In the Netherlands, there are three different companies which measure these roads: KIWA-KOAC, Aveco De Bondt and GRiP Road Inspection. The measurements are harmonized at European level under the responsibility of the BASt, the German variant of Rijkswaterstaat. Previous research has been done by Q-consult progress partners who is hired by RWS to formulate a model. See Table 1-1 for a short overview of the organizations.

Organization	Role	Note
Q-consult PP (and David)	Researcher	Q-consult PP is hired by RWS to investigate the influences and make a model.
RWS	Road owner	RWS is responsible for the quality of the roads.
KIWA-KOAC, Aveco De Bondt and GRiP Road Inspection	Executes measurements	Hired by RWS to perform the measurements needed to evaluate the road surface. They use the SWF method for measuring.
BASt	controller	Controls if the quality satisfies the European norms.

Table 1-1 Stakeholders for the model

1.4.2. Literature review

In this part, more information about the skid resistance research and statistical analysis will be done. This information will help to identify possible influences according to physics and how to determine the influence. The first part of this review is to identify potential influences and how they influence the skid resistance. Since otherwise, there are endless possibilities of expressing variables and determining their influence. In the second part, the research is focused on helping to identify important variables and formulating a model. This part is about the indicators, these tell us something about a variable, but also how the data should be used.

1.4.3. Gathering data

The data needed for the amount of rain during the measurements can be found in the KNMI database[7]. For this research, some elements (found in the literature review) are selected that might be interesting for the skid resistance. After retrieving this data, it will be used to answer the research questions. The measured data is available for Q-consult and me to use during this research. This data is retrieved by the road inspectors.

1.4.4. Analyzing data

After all the data is collected and ordered, the data can be prepared for analyzation. For this part, it is easy to use a program like Minitab, which they do have at Q-Consult PP, but for this report and with the given time Minitab will be used. With the use of Minitab multiple regression analysis, we can determine the influences a formulate a model. Based on this analysis, the influences of the variables can be determined.

1.4.5. Making the model

In this stage of the report, the influences of the additional variables are determined. With this information, a choice can be made for the kind of input. There are multiple options to measure these variables. Rain in the last 24 hours can, for example, be expressed in ml, but another option is to use it as yes or no (binary). In this stage, it is important to keep the usability of the formula in mind. It should not be too complex to use and find the data needed for the correction. Therefore, some trade-offs can be made, and multiple formulas can be formed.

1.4.6. Choosing the formula

This is the last stage. In the last stage, multiple models are formed. Here the best model can be chosen based on its accuracy, complexity and other aspects which will be later determined. Based on this formula, the right restrictions can be added to increase its accuracy further.

1.5. Project scope

The goal of the assignment is to make a model that can correct the measured SWF value to the actual SWF value. The addition to this model is the influence of rain and drought.

For this model, it is important to know the influence of these factors on the short-term of the measurements. The long-term influence must be treated differently since their influence should be included in the measurement. This is already done by the current situation, but there is still noise remaining. To formulate an improved model, we will focus on the influence of rain and drought and how it can be measured. Research has already been done to determine the influence of temperature and seasonal deviation. We will take the old model into account for this part but expect that their values will also change since there might be a correlation between all these variables. This means that our model will also contain the variables of the temperature and day since they were already included in the old model, but their influence will be different. These are the primary variables which will be evaluated, but a small research to other variables will be added.

For this report, the measurements of the new research are not yet done. Thus, the model will be based on the values found in the old model. However, my goal is to obtain our model in a way that can easily be replicated with new data. A large part of the work will be to collect and retrieve the data of rain on the measured days. This data is different for every measurement set since they are all in different locations and different days. We, therefore, limit this research to the influence of rain and drought, these can be expressed in many different variables, and we first have to find the right expression.

1.6. Deliverables

The goal of this research is to formulate a new model, which can calculate the real value accurately with a small standard deviation. To complete this, I must start by retrieving the data needed for my regression model. Based on the new data, multiple new independent input variables can be formulated. The influence of these variables should be calculated independently and dependently of the other input variables. These findings are used to make multiple models using different methods. These models can be evaluated and help to choose the best model(s). I expect that every model has its own weaknesses and strength. For example, some can be very accurate, but too complex or have a lot of restrictions. Thus, I will write a report with my evaluation for these models describing their advantages and disadvantages. This can be helpful since the preference of Rijkswaterstaat can change or when new research starts with other data.

1.6.1. Report

The largest deliverable will be the report. In this report, every choice I make is explained and what I did. The goal is to explain the choices made and show the working method. In the report, other deliverables will also be provided. This will include examples for the choices of variables, arguments for the best model and other choices. Therefore, it contains part of the other deliverables.

1.6.2. Weather data

For the regression analysis of the weather conditions, the KNMI database is used. For the analysis, the data will be transformed to compare the same variable using different units. This will be provided for RWS and QCPP to show on what the analysis is based. This data format will be in either excel or python since these programs can quickly transform and order data.

1.6.3. Improved Model

The goal of the research is to improve the model. This new model is the deliverable for QCPP and RWS. Most likely will there be multiple models since there might be different best models in the perspective of each stakeholder. The models exist out of the formula to correct the SWF value and the requirements and restrictions during the measurements.

In the final document the “best” model will be given with the other models. The reasoning is included based on the found data and substantiated by figures. The models can be tested and compared to the old model by using data to retrieve corrected values. In this part, the standard deviation and accuracy will be compared to see if the new model is better than the old model and what could be further investigated. We expect that this model has higher accuracy if rain is used as an input variable. We also confirm whether the drought restriction is enough or should be changed.

1.7. Methodology CRISP-DM

For this project, a methodology specifically for data mining is chosen that helps us understand and transform data. For this report, the goal is to improve and test a model that is based on the data. Therefore, the data mining methodology is important in this report. Since data mining helps to process and discover patterns in datasets. The model, in this case, is a formula with restrictions and requirements correcting the SWF of a measured road to the expected SWF during standard circumstances.

For this project, the CRISP-DM methodology is used; this is proven to be a flexible tool helpful for data mining[8]. This methodology consists of 6 steps which tasks can be performed in different orders. The first 5 stages will be implemented in this report; step 6 is the conclusion with recommendation since this step mostly is implementing the model.

1. Business objectives
2. Understanding the data
3. Preparation of data
4. Modelling
5. Evaluation
6. Setting out

The stages are implemented over multiple parts in the report[6]. In the following part, the meaning and importance of each step are explained.

1.7.1. Business objectives

The first stage is to understand what the goal is from the business perspectives. In this stage, important factors that influence the outcome are determined. To fully understand the goal, the following questions, need to be answered; what the desired outputs are, what is the current situation, and what should be the data mining goal. Here the desired output is the main goal, and the data mining goal(s) expresses this goal in technical terms. For this project, the business goal is to improve the model that corrects the measured SWF data. Therefore, the data mining goal is to determine the influence of independent variables to increase the accuracy of the model.

1.7.2. Data understanding

The second stage is to acquire the data which is stated in Chapter 2. This includes the data and methods used to understand and analyse the data. Literature and data analysis methods are explained here and used for the data understanding in Chapter 4.

1.7.3. Data preparation

The third stage is data preparation. In this stage, the data is observed, and restrictions are made. The goal is to decide which data is excluded, which data is cleaned and the transformation of data. In the previous stage the data is already analyzed individually to understand it; in this stage, the data is prepared for modelling. This is included in Chapter 4. Here the data is transformed, and a selection of predictors is recommended based on the determined influence. This chapter also include the data transformation, this is necessary since some data must be expressed in another unit or excluded for an analysis.

1.7.4. Modelling

The fourth stage is to model the data. This includes the way the model is chosen, the assumptions that are made and the built model. Here the data retrieved from the preparation step is used for a multilinear regression analysis to formulate a correction formula. This stage is performed in Chapter 5. Here multiple models are given with a summarized list of there qualities on which they are assessed. In this chapter, the restriction for measuring after a period of drought is analyzed. This last phase is important since it can improve the accuracy of the model a lot if drought has an unexpected or unpredictable influence on the SWF.

1.7.5. Evaluation

The fifth and last stage, which is included in this report is the evaluation of the model. This step is included at the end of Chapter 6. Here a model is recommended with an explanation including restrictions and other aspects. Then this model and the old model are used on a large dataset to calculate their accuracy again and determine the improvement.

2. Literature research

In this chapter, the literature which is relevant for this report will be addressed. The literature can be divided into two parts. The first part is about previous research regarding skid resistance. The focus of this part is to find and confirm possible influences on the skid resistance. This helps to answer a part of the first sub-questions. The second part is about the calculations that are used in this report. This helps to determine the influence of variables and with answering research question 3 and 4.

2.1. Previous research regarding possible influences

In this part, previous research is used to find possible influences on the skid resistance and find more information about their relation. This helps to solve which variables are possible influences on the skid resistance and how they should be expressed. Expression means that for instance that rain can be expressed as an amount in the last four days or the last time since the rain has occurred.

2.1.1. Influence of temperature

The current model of Rijskwaterstaat[2] (see Chapter 1.1.3), already uses temperature as an input variable for the correction. The current model uses the temperature of the water and the temperature of the road surface for the correction. Other research reports also confirm that skid resistance is strongly influenced by temperature[4]. During the measurements of the SWF, the temperatures of **air**, **water**, **road surface** and **tyre** were also measured. Therefore, the calculations are limited to the influence of these temperatures.

In the research of Ed Baron [4] the influence of temperature was investigated on the skid resistance. In addition, research was done to determine which temperature were independent influences on the SWF (the cause) and which were just correlated with the friction (influenced by the cause). The research concluded that there was not a unique relation between air and road surface temperature. In addition, the most direct influence of temperature on the skid resistance appeared to be the road temperature. This, however, has not yet been confirmed with the addition of the temperature of spray seals (water) as a relation.

Another research found a different relation between the tyre temperature and the SWF coefficient. Instead of linear, it would be $Theta1 + \frac{Theta2}{(Theta3 + TempTyre)}$ [9], where Theta 1, 2 and 3 are 0.63, 45.9 and 80, respectively. Therefore, a non-linear test will be performed. This research was, however based on one road surface; therefore, the Theta1, which is the baseline, cannot be calculated precisely since the roads in this research have different standard values.

2.1.2. Influence of rain and drought

In a publication sponsored by the Committee on Surface Properties Vehicle Interaction[10], the short term effect of the rain was researched. Here the relation between rainfall in the prior days before measuring was researched. The research concluded that there is an effect of rain on the skid resistance. The skid resistance would decrease during dry periods and increase after heavy rain. However, the correlations coefficient between the skid resistance and rainfall, expressed in WRF (weighted rain function), were found to be consistently low. The same relation between skid resistance and drought showed a significant improvement in the correlation coefficient.

In the current model, rain is not an input variable. It has been suggested that rain does influence the SWF, but the occurrence itself and not the amount of rain. Therefore, Rijkswaterstaat requested to look more into the effect of drought (periods without rain), since this already is a restriction.

2.1.3. Influence of seasonal variation

The current model of RWS uses seasonal deviation as one of the input variables to calculate. Seasonal variation is a broad term for all kinds of variables together, which each have an influence on the SWF value based on the day or week or month. The current model of RWS uses day number as an input variable to correct the SWF measurements. In previous research between the seasonal deviation and skid resistance, a similar correlation was found[11]. The seasonal effect would have the shape of a sinusoid with the period of a year.

2.2. Regression analysis

In this part, research is done to find how we can analyse the influence of variables and which calculations are relevant and could be performed. This helps *to solve how influences should be determined, what makes a model complex,* and *how to choose the best model.*

2.2.1. Choosing best predictors

In multiple linear regression, not one predictor is determined, but two or more. If multiple predictors and their coefficient are determined at the same time their influence, the coefficient is determined (more) independent of each other. After doing such an analysis with multiple predictors, the question remains, which are the best predictors. Therefore, we start off with important indicators.

Pearson correlation
The Pearson correlation coefficient is commonly used to measure how strong two variables are correlated. The coefficient is a number between +1 and -1 which indicates if there is a positive or negative linear relation and its strength. A coefficient of 0 indicates that there is not a linear relation found. [12]

P-value of the variable
When you perform a statistical test, a p-value helps you determine the significance of your results in relation to the null hypothesis. The null hypothesis in our case says that the variables are not correlated. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis[13]. When calculating the influence of one predictor, a T-test is used to find this value, in multiple linear regression, the F-test is used. This does not mean that P-value necessarily indicates if a variable is practically important.

R-Squared
R-squared represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. After multiple linear regression analysis, the R-squared can be calculated. The R-squared in our analysis is the **percentage** of values which are on or **close enough to the expected SWF value** after correcting it (see Assumption 0). Therefore, the R-squared of each predictor individually or the added R-squared in multiple linear regression indicates their added relevance for the calculation. The greater the increase in R-square, the more relevant the addition of this predictor is.

Standardize regression coefficients
Each predictor has its own coefficient; this coefficient is used for the eventual model and helps to correct the SWF. However, since the predictors can have different scales and units, it is not possible to compare them directly. A standardized regression coefficient can be compared since they are recalculated to the same scale. [14]

Suggested influencing factors
In this report, influences are determined using mostly regression analysis. If a predictor is similar or influenced by another predictor or a lot of possible variables are analyzed, the predictor might not be the best or even a good option. Therefore, in the previous part, the results of research of possible influences are done to find suitable predictors. This does not mean that every predictor most have a physical influence, as temperature has on friction, for example. Other predictors like day number, which indicates the seasonal influences can still be used. But the variable should make sense.

2.2.2. Choosing the best model

Choosing a suitable model is like choosing the best predictors since a model exists out of the best predictors. However, in this part, we no longer want to determine relevant predictors, but a not too complex and accurate model. For a model, the choice of which predictors are relevant, is the same as choosing a predictor. The next part is to deselect or reselect some variables.

The variance inflation factor (VIF)
The VIF detects the multicollinearity in a model. The multicollinearity is the correlation between predictors. High multicollinearity between predictors is unwanted since it makes it harder to determine the individual influence of a variable. In most cases, a high number of variables leads to higher VIF values. The influence is calculated through regression, not physics, and bases the coefficients on improving the R-squared, even if it does not make sense.

$$VIF = \frac{1}{1 - R^2}$$

The R-squared value is calculated by regressing a predictor against every other predictor in the model. Higher VIF values indicate higher correlated predictors.[15]

Complexity
An important aspect of each model is its complexity. With complexity, the simplicity of the model and ease of usage is meant. Therefore, a model with a high R-squared value might still be less reliable and usable than a simpler model. This, for example, can be the case if a model uses a variable which is hard to measure. Another problem can occur when a model is to simplify and only uses predictors that indicate the possible circumstances.

The complexity of input variables is subjective; therefore, the use of these variables has been discussed with the stakeholders. For example, it takes a lot of work to measure the exact amount of rain in the previous days on the road. Thus, is the use of rain expressed in the exact amount a complex variable.

2.2.3. Data usage

An important aspect for this report is the data usage. An option is to not use all the data but only a percentage and use the other part to test the models on. The prediction analysis we use to determine the influences is a part of machine learning. In our experience with machine learning, you should always use all data you have to get the best model. Of course, this does not leave any data to control, as some studies do[16]. However, since we do not know the actual SWF value, we perform calculations without a control group, we cannot test the model with certainty.

3. Assumptions

For this research, some assumptions had to be made to be able to perform a regression analysis. Some were already made in previous research. In this chapter, the assumptions are explained.

3.1. The expected SWF value

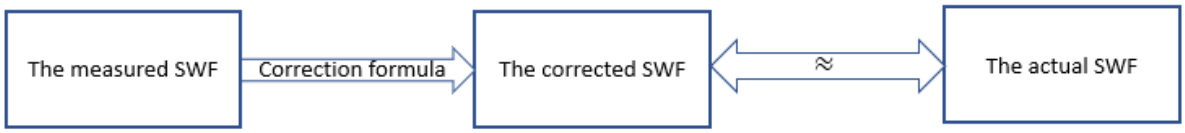

Figure 3-1 Representation of the outcome of the correction model and the needed value

The first assumption is one of the most important ones for this research and is also used in the previous research for RWS. Since it is nearly impossible to know the actual SWF value, it is not possible to calculate the influence of the variables based on this difference. Therefore, a standard circumstance has been defined as a measurement performed on 15 June when all the temperatures are 20 degrees, and 1 mm of rain has fallen the day before measuring.

In this report, we want to correct all the values to the expected SWF at these circumstances. **Therefore, we assume that when the measured SWF value is corrected, this value is nearly the same as the value during these circumstances.** The influence of these factors is calculated based on measured SWF values during other circumstances. A problem with this assumption is that different models can have different actual SWF values, more about this later.

3.2. Difference between variable and result

This assumption is strongly related to the first one. The influence of these factors is calculated based on measured SWF values during other circumstances. Thus, **we assume that the correlation between the measured SWF during other circumstances and the difference in these circumstances can be used for the correction formula**. After using this correction formula, the corrected SWF is nearly the same as the actual SWF.

The coefficient of the influencing variables used for the calculations are based on the difference between every measured SWF and the mean of the expected corrected SWF value. Again, we remain with the problem that the different models can have different actual SWF values.

3.3. The accuracy of a model

In the first assumption, we concluded that the SWF value corrected by different models gives different results. Therefore, we would have multiple actual SWF values. Thus, our accuracy only represents the number of measurements that can be explained using that specific model, all having their own expected actual SWF value.

Since we cannot know the SWF value, we made assumption 1. However, we are now left with possibly multiple actual SWF values. Thus, **we assume that high accuracy in a model indicates a higher chance of correcting the actual value, if the model is not overfitted and the used variables are relevant.** Relevant variables are significantly proven correlated with the measured SWF. Overfitting is using too many variables for the model, which always leads to an increase inaccuracy. Therefore, the chose for the right actual SWF value depends on the trade-off between the accuracy and complexity of the model.

3.4. Rainfall near measurement places

This assumption is about the area of rain and how local it is. Since the exact amount of rain is not measured for each measurement place, only KNMI neerslagstations[7] can be used. There are 325 stations that track the amount of rain each day. Therefore, we can retrieve the amount of rain of 325 spots in the Netherlands.

Since rain can be very local, it is hard to find the precise amount of a measurement place. In addition, is every measurement based on a 2 km long road, and therefore the amount of rainfall may also vary within a measurement. To still do some calculations with the rain we have to assume about the amount of rain at a station and at the measurement place. **We assume that the amount of rain at a station is the same at the measurement place if the mean is within 4 km**. This assumption was agreed upon by the stakeholders since it is hard to determine the area of rain.

3.5. The decrease in SWF

In the introduction of this report, the average decrease in SWF is already discussed. It is expected that the SWF on average decreases with about 0,02 a year. The decrease in SWF is higher for new roads or heavily used roads. Since these influences have a permanent effect on the SWF and are not short-term effects, it should not be used for the correction model. Since the measurements are performed over a period of two and a half year, the actual SWF should have decreased of this period. Therefore, **we assume that the expected SWF of a measurement place is the same in one year**.

3.6. Same seasonal influence in the Netherlands

One of the variables for the correction is the day number. The day number is a predictor for the seasonal influence on that day. The seasonal influence changes over the year as has been demonstrated in the introduction. For this research, **we assume that the seasonal influence does not variate within the Netherlands**.

4. Which variables should be included in the correction model?

The skid resistance in the Netherlands is measured using the SWF method[2]. The measured skid resistance is a snapshot influenced by specific circumstances. During these measurements, water is added, to simulate a wet road, this should negatively affect the friction condition according to RWS. The results of the measurements from different locations and dates cannot be compared directly since the circumstances during the measurements are different. Since it is hard to recreate the same circumstances for each measurement, a model needs to be made to correct the measured values to the expected value during standard circumstances. The goal of this chapter is to find possible input variables for the correction. To see which independently influence the skid resistance. Then to determine their influence on the skid resistance and transform them to input variables for the model.

In this part, the analysis is based on the correlation between a possible influence and the measured value. This means that the variable in the correction formula only is a predictor, useful for the correction, but it does not have to be the direct influence. Choosing the cause could. A proxy variable that is simply correlated to the response and is easier to obtain than a causally connected variable might produce adequate predictions. An example can be the influence of the air temperature outside that is considered in the seasonal deviation (this might be expressed in the day the measurement took place). For every analysis, the year and measurement place are taken as categorical variables since the actual value is different for each place and should be lower each year. This makes a different group for every measurement in place and year, e.g. location x in year z.

The evaluation of the variables is based on their relevance and accuracy. Relevance can be expressed in the individual and additional accuracy of a variable, expressed in the P-value of a variable. The accuracy can be expressed by the R-squared of a model that uses a variable.

The coefficients representing the influences of variables are less important. Their size in the formula depends on their scale. For instance, if the analysis is done using a variable expressed in millimetre, the coefficient is much smaller than the coefficient expressed in a metre. Since every influence has a different unit, it is hard to compare them. Therefore, the variables are evaluated on their relevance instead. This method uses the P-value, which either rejects the null hypotheses (null hypotheses states that there is no correlation) or fails to reject the null hypothesis. To find the P-value, a T-test is performed.

4.1. Possible influences

The variables which are important for the correction are those who have a short-term influence on the skid resistance. For this report the short-term influences are defined as reversable changes in the SWF of the road. Influences which permanently change the SWF are not short-term influences since this also influences the SWF during “standard circumstances”. The long-term influences, such as cracks in the surface, must not be corrected out of the data, since these influences are present during normal road-usage. The influence of short-term effects that only change the value during or till short after these circumstances took place, are not taken in account. Therefore, only circumstantial influences with short-term influences are determined in this section. In the current correction the influence of temperature and as a restriction drought is used by Rijkswaterstaat.

4.2. Temperature

The goal of the new, as well as old, correction model is to correct the circumstances to a standard situation (when the temperature is 20°C). This agrees with a report of the international surface

friction conference of 2011[3]. Therefore, the correction is based on the difference in temperature between the measured situations and the standard situation. In the current model, it is the difference between the water temperature and the road surface temperature. With this data a regression analysis is done, the results show the relation between the temperatures and the SWF is. This is the regression of four temperatures on the SWF with year and measurement place as categorical variables.

4.2.1. Data preparation

In the following part, calculations are made to find the influence and importance of each temperature. For these calculations, all data can be used except some outliers which have been identified. During the calculations, a disturbance in the influence of tyre temperature was noticed. Therefore, tyre temperature has been excluded from further calculations after this was identified and discussed.

4.2.2. Results influence temperature

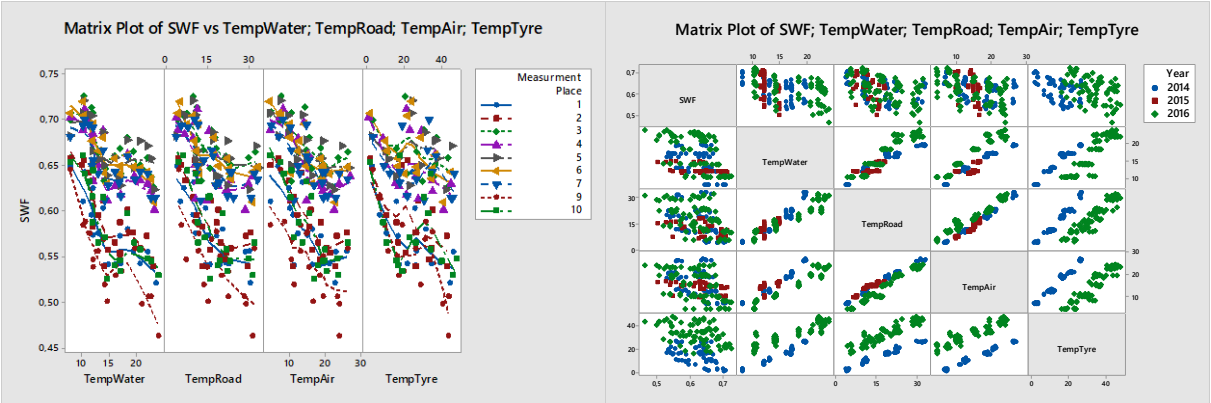

Figure 4-1 SWF vs. Temperature

In Figure 4-1, the temperatures are expressed in Celsius during measurement and the SWF as a coefficient. In the left figure, the colours represent different measurement places, and the right figure represents the year. In the graphs with water, road and air temperature there is a negative relation between the temperature and the SWF.

The temperature of the tyre also has a negative effect on the SWF, but this seems to be inconsistent. In the left graph, all the relations between the variables are displayed with each other. It becomes very clear that the temperatures are strongly interdependent. The relationships with the tyre temperature also show two different correlations in one relation between the other temperatures that differ per year. In addition, the temperature of the tyres only includes the year 2016 and 2015, therefore, it contains fewer data.

In Table 4-5, the influence of each temperature individually is determined on the SWF. The Pearson correlation coefficient shows the strength of the correlation, which shows **a medium** and negative linear **correlation** for **water-, road-** and **air** temperature and a **small** and negative linear **correlation** for the **tyre** temperature.

Correlation	SWF	TempWater	TempRoad	TempAir
TempWater Pearson correlation	-0.450			
TempWater P-Value	0.000			
TempRoad Pearson correlation	-0.423	0.918		
TempRoad P-value	0.000	0.000		
TempAir Pearson correlation	-0.452	0.854	0.939	
TempAir P-value	0.000	0.000	0.000	
TempTyre Pearson correlation	-0.275	0.787	0.688	0.504
TempTyre p-value	0.001	0.000	0.000	0.000

Table 4-1 The correlation between SWF and Temperature

The weaker correlation between the SWF and tyre temperature and the odd representation in Figure 4-1 has been discussed with RWS. It turned out that in 2016 another company performed the measurements than in 2014. The SWF method is the same for both companies; however, the way of measuring the tyre temperature was not included. Therefore, the different correlations between the same two variables can be explained if the method of measuring tyre temperature differs. The P-value is low for every variable; therefore, the chance of a correlation between the temperatures and SWF value is significant (Temperature has an influence).[17]

The same table also shows a strong correlation between every temperature individually. This indicates that the temperatures influence each other significantly, which makes it hard to determine the influence of the temperatures independently in a multilinear regression model. This correlation agrees with the strong relations in Figure 4-1.

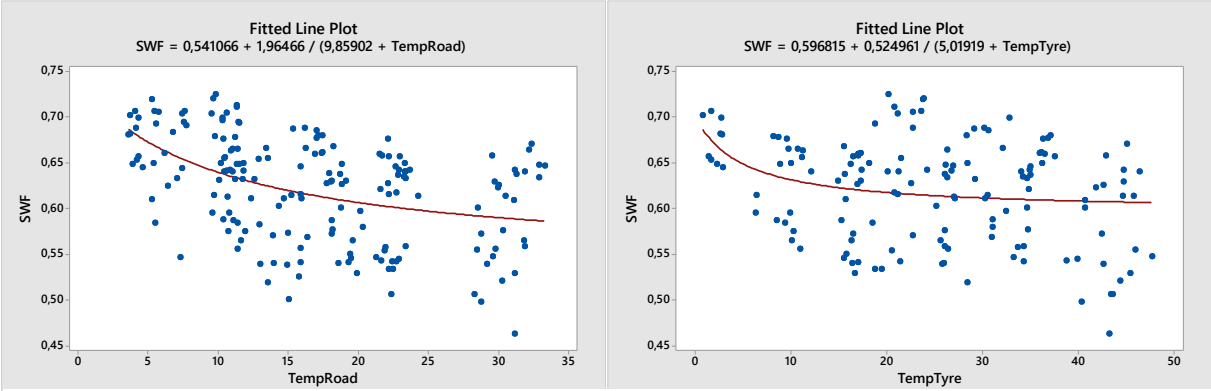

Figure 4-2 Non-linear Relation Temperature and SWF

The last test was done with the non-linear relation found in research of the University of California[9]. The result of this relation can be seen in Figure 4-2, where the relation is shown for the temperature of the tyre and road surface. The results also show an S-value, the standard error of the regression, of 0,0541347 and 0,0504191, respectively. This value gives more information about the quality of the fit; however, its result is less meaningful than that of R-square, that is used in linear models. Because of the relation between the temperatures themselves found earlier, the road surface temperature is also used, which seems to have a better result than the tyre relation.

4.2.3. Conclusion influence temperature

The P-value tests show that all the temperatures have a significant correlation with the SWF value. In addition, all the variables have a negative influence on the measured SWF value and should, therefore, be corrected positively in the correction model. However, from previous researches and as shown in f Figure 4-1 a strong relation between each temperature is found. This relation makes it

hard to determine the precise influence of each temperature variable independently in a multilinear regression model. Therefore, the model should not contain all the temperatures as input variables since they are all correlated. Only one should be enough. But if it includes multiple temperatures, they should all have a positive coefficient.

Based on the data, the tyre temperature should be the first to exclude, since it has a gap in Figure 4-1 between the years and contains fewer data. Road surface temperature should be included based on the literature. There might also be a non-linear relation; the quality of this relationship is, however, hard to determine. There are still a lot of other variables influencing the outcome. The measurement method of temperatures is also important since it influences the correction because it differs per vehicle or organisation. Therefore, water could be the most stable to measure and use from all the temperatures; road surface might have the most influence on the SWF.

4.3. Rain and Drought

The currently used method to measure the skid resistance uses the addition of water while measuring to simulate rain. This is the standard method since rain lowers the skid resistance, which gives the SWF value as a result under a bad short-term condition. However, there are some differences between this simulation of rain and actual rain. One difference is that the amount of rainfall can differ, while for the measurements, the same amount of rain is used. Another difference is that the water of rain spends more time on the road, while in the simulation, the sprayed water is immediately followed by the measuring tyre. This might give the water time to fill the small cracks in the surface, which can have an influence. The last difference that might be noticeable is the composition of rainwater and of the water used in the machine that uses ditch water.

4.3.1. Data preparation

The measurements do not include data about the amount of rain. Therefore, these additional data need to be gathered in another way. In the Netherlands, there are two types of “stations” which track information about rainfall: weather stations and rain stations. Weather stations measure and store a lot of information detailly expressed, while rain stations only store the amount of rain and snow in the last 24 hours. However, there are more rain stations (325) than weather stations (34). The stored data of both stations are place-specific, while the data of the road measurements is based on a 2 km long surface. Thus, in this part, only the measurement places within 2 km of weather- or rain station are used.

After including the amount of rain which has occurred the days before measuring, multiple transformations have been made. The rain has been expressed as a total amount in the last X days before measuring, days since at least a certain amount of rain has occurred.

4.3.2. Results influence rain

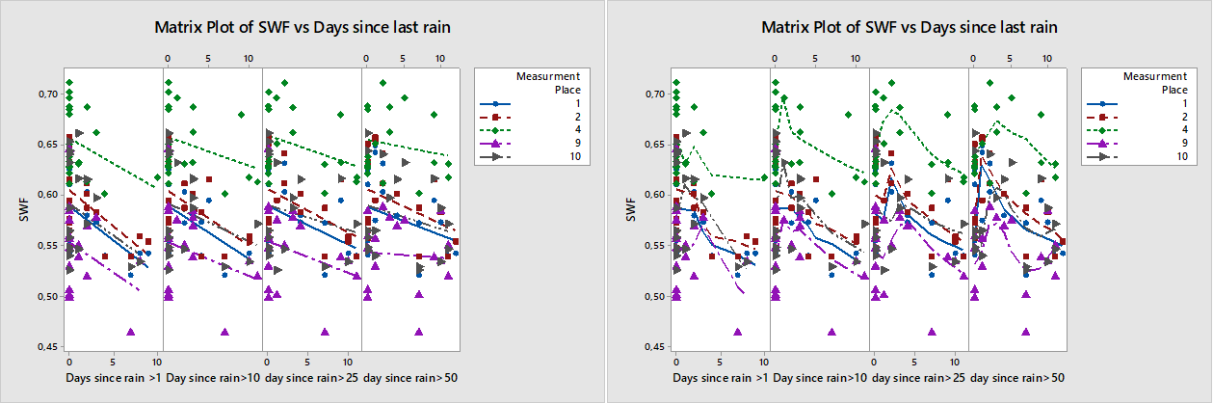

Figure 4-3 SWF vs. Days since last rain

For this part of the research, data is gathered from nearby rain stations and sorted to amount of rainfall in the days before measuring. The data contained the amount of rain of every day per 0.1 mm. The days were in 24 hours intervals from 08:00 UTC in the day before till 08:00 UTC that day. Previous researches suggest that this data should be transformed since the amount per day is not as important as the last occurrence itself[18]. Therefore, the first transformation is to ordinal data by ranking the data to days ago since more than X rain has occurred (here X is amounts of 0.1mm). For instance, X is 10 the second graph should be taken, this shows the relation between the measured SWF and the maximum amount of days since 1 mm or rain within one interval has occurred.

According to previous research, the measured SWF should be lower after a long time without rain since the roads will be “dirty” which has a negative effect. This has been done for rainfall per day with a minimum amount of 1, 10, 25 and 50 for X (which is in mm per day).

In Figure 4-3 the last time since the 0.1-, 1-, 2.5- or 5-mm rain fall can be seen in relation to the SWF coefficient. Based on this and the theory, a linear regression analysis is performed with rain as a categorical predictor instead of continuous, since theory expects an upper limit for the decrease. For all groups, significant influence is found since P-value is 0.001.

The next goal is to determine if there is a limit or if the rain occurrence should be used as a binary. For this analysis, the data is transformed from a 2,5 mm column. This data is split into a group where the maximum amount of days since rain is 7; thus, after seven days, it stops searching. The other groups are binary and show a one if it has not rained in the last Y days, where $Y = 1, 2, 3, 4, 5, 6, 7$.

Correlations SWF	*Days without rain (max 7)*	*1 day no rain*	*2 days no rain*	*3 days no rain*	*4 days no rain*	*5 days no rain*	*6 days no rain*	*7 days no rain*
Pearson Correlation	-0.207	-0.098	-0.166	-0.237	-0.244	-0.244	-0.244	-0.14
P-value	0.039	0.33	0.098	0.017	0.014	0.014	0.014	0.165

Table 4-2 Correlation SWF and Rain

In Table 4-2 the results from the analysis of rain as an influence on the SWF is shown. From this table, it is directly visible that no difference could be determined between 4, 5 and 6 days without rain before measuring. The reason for this is that there was no data difference in these groups. Meaning that the distribution of data where more than X amount of rain has occurred is the same for these days. Based on this finding, the best choice would be to take either 4 or 5 days without rain since previous research[10] shows an upper limit on day 7.

4.3.3. Conclusion Influence rain and drought

Based on the correlation analysis and the results of the p-value, there is a significant correlation between SWF and rain and drought. However, this relation does not appear to be linear or can be expressed in amounts based on the results and literature. Therefore, the influence of rain could be used in the correction model in two ways. One option is to include it with an upper limit; theory suggests seven days. Another option is to include its dummy variable by including if there was that 2,5 mm rain a day in the last X days. Rain is not used in the current model; in the new model, rain could be used if the accuracy outweighs the complexity.

The dummy or upper limit variable influences the cleanness of the road.[2] The influence of a dirty road has been taken into account by restricting measurements if no more than 1 mm rain occurred in a 24-hour interval in the last 14 days. Therefore, rain always makes the model complex since it is hard to measure.

4.4. Seasonal effect

The seasonal effect depends on the day number, week number or month. In this part, these options will be researched and evaluated. Starting with the currently used factor, seasonal effect expressed in day number as a sinusoid function. After that week number and month will be expressed as a sinusoid. After expressing them as a sinusoid, they will be researched as categorical variables. This will give an addition or subtraction per time interval, which is a standard for all the values within this interval.

In this chapter, the correlation between seasonal variation and all the other variables are tested, individual. Which can have a large effect on seasonal variation since temperature and might also be strongly correlated with seasonal effects, like temperature. Therefore, this will also be researched to see their correlation.

4.4.1. Result influence seasonal variation

For the tests with the seasonal variation, the dates are transformed to multiple different kinds of variables. The dates are transformed to day number, week number and month number. Measurements are not performed in January, February and march since frost have a totally different effect on the slip resistance. The relation between the SWF and the time period can be seen in Figure 4-4A pattern in this relation is visible. The right figure shows the deviation from the function per measurement place; this shows that the SWF during normal circumstances is different for each measurement place. This is expected since the quality of the road differs per place.

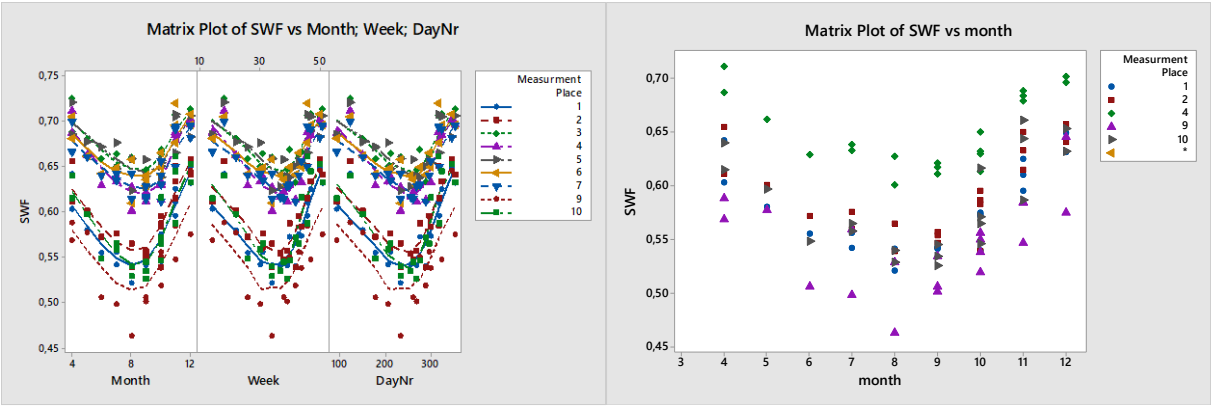

Figure 4-4 SWF vs. Time period

The pattern in graph Figure 4-4 could be a quadratic (if a one-year interval is used) or sinusoid function in which case an infinite interval can be taken. Since the relation is in an interval of a year and based on the literature, this most likely is a sinusoid. The next step is to determine a sinusoid which covers this data. The period of the sinusoid should cover a year; the other variables should be the same. The vertical shift is the only variable that should be different for every measurement place since they are in a different condition.

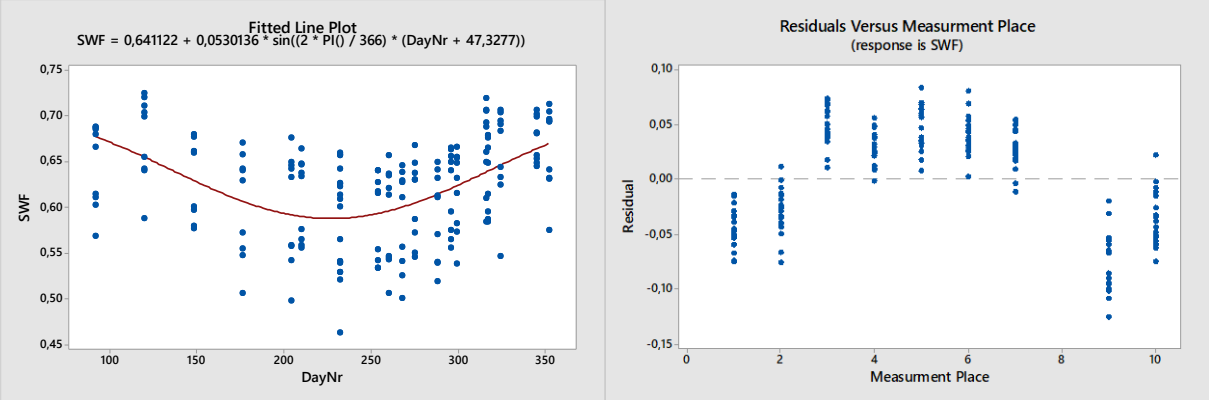

Figure 4-5 Non-linear function SWF vs. Seasonal variation

In Figure 4-5 the found sinusoid can be seen. The other graph shows the remaining residual from the current generally corrected SWF value vs the measurement place. This figure is interesting since it shows that the remaining residual is strongly correlated with the measurement place, and thus varies based on the road condition. This indicates that each measurement place is lower or higher than the corrected value by this function. Part of the remaining difference varies on the quality of the road. Therefore, it is lower for measurement place 9, which probably has a worse quality than measurement place 5.

$$DifferenceDate = sin\left(\frac{2\pi}{Time\ periods\ a\ year}((measurement\ date) + 0.129 * Time\ periods\ a\ year)\right)$$ *In an Equation 4-1 Time period calculated to sinusoid Date*

Additional correlation analysis with all the time intervals vs the SWF value, no linear relation was found since all the P-values were higher than 0.3. In the literature and previous research, a relation was already determined. This seasonal variation is predicted by using day number as an indicator, which has the influence of a sinusoid.

For a sinusoid, we already know the timer period since it varies over the course of a year. For the period $\frac{2\pi}{Time\ periods\ a\ year}$ should be used. This tells us when the sinusoid should be back at the starting point. Here *time periods a year (TPAY),* is dependent on the unit which is used to express the X (measurement date). This should be 12 for months, 52 for weeks or 365 for days.

For the next part, the measurement date is the variable in the sinusoid. This changes the outcome of the sinusoid and calculates the intensity for seasonal influence for each day. The 0.129 is derived from the analysis done in Figure 4-5, and represents where the sinusoid starts. This is a date which represents when the influence should by seasonal variation should be 0. See Table 4-3 for an example.

Expression	TPAY	Date	X	0,129*TPAY	$\frac{2\pi}{Time\ periods\ a\ year}$	Y	Sin(Y)
Days	365	4-feb	35	47.085	0.0172	1.946	0.93
Weeks	52	8-aug	32	6.708	0.1208	4.677	-0.9994

Table 4-3 Example calculation difference date

After correcting the measurement dates with this sinusoid, a linear regression can be performed since the date now are transformed to a linear scale. The date values recalculated to the difference they have on a sinusoidal scale from day 47. The corrected values should now all have a value which varies between -1 and 1. Now we can calculate how the day number on a sinusoidal scale is correlated with all other variables. The results are shown in Figure 4-6 and Table 4-4.

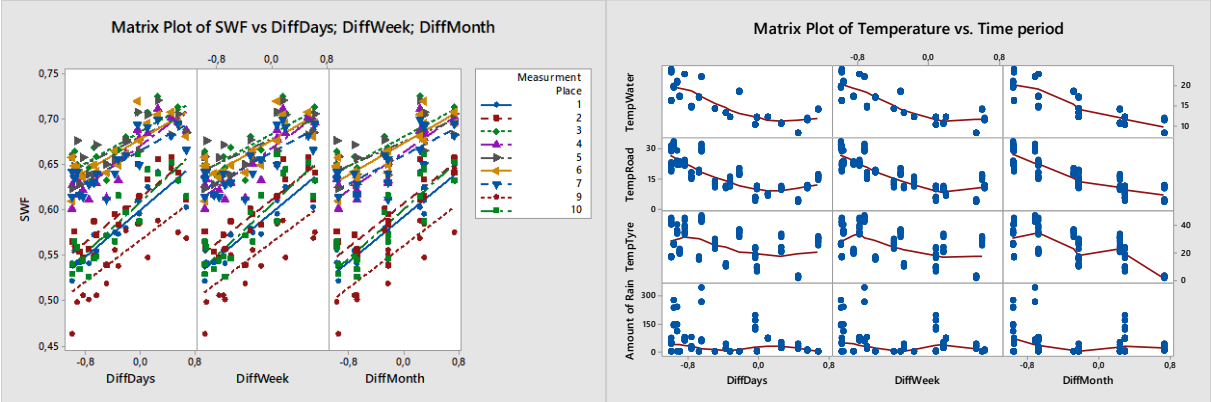

Figure 4-6 Relations sinusoid Date

Correlations DifferenceDate	*SWF*	*Temp Water*	*Temp Road*	*Temp Tyre*	*Amount of rain*	*Days since last rain >25mm*	*4 days <25mm of rain*
***Day** Pearson Correlation*	0.505	-0.754	-0.738	-0.379	-0.245	-0.193	-0.321
***Day** P-value*	0.000	0.000	0.000	0.000	0.001	0.055	0.001
***Week** Pearson Correlation*	0.499	-0.803	-0.794	-0.458	-0.271	-0.162	-0.283
***Week** P-value*	0.000	0.000	0.000	0.000	0.000	0.107	0.004
***Month** Pearson Correlation*	0.501	-0.836	-0.850	-0.583	-0.279	-0.126	-0.217
***Month** P-value*	0.000	0.000	0.000	0.000	0.000	0.212	0.030

Table 4-4 Correlation sinusoid Date

The relation between the sinusoid of date and SWF is visible in Figure 4-6 and Table 4-4. The P-values in the table show there now is a significant correlation between the date and most the relations. These relations were not significant for the others calculation. The time interval before the sinusoid correction (the lowest P-value was >0.3), and therefore no linear relation was proven, a sinusoidal correlation is.

If we take the standard significance level of $\alpha = 0.05$ all the time intervals are significantly correlated except for the variable *Days since last rain*. The calculations with the day as time interval give the highest test results with the lowest p-values. This could be expected if the function of SWF is sinusoidally correlated with the day number since it has the narrowest interval and fewer outliers from this interval. Another reason for this difference can be that the correlation calculation is based on all the measurement places seen as the same group.

We, of course, notice a very high negative correlation between the temperatures and date expressed in a sinusoid. This means that if the difference date is high in value (highest around 13-Feb), the temperatures are lower. Therefore, the highly correlated Pearson correlation coefficients make a lot of sense. This suggests that seasonal variation is a very good predictor for the temperatures. However, since we already know the temperatures and direct influence by temperature on the skid resistance is suggested, using temperature in the model helps to correct even more accurately.

Rain is also corrected by the sinusoid of date; however, the Pearson correlation is much lower for rain amount. Seasonal variation does slightly indicate if it has rained in the past 4 days.

The results from the calculation of the coefficient and significance for every time interval individually with measurement place as a categorical variable can be seen in Table 4-5. The categorical coefficients are only shown for these months since there is only enough data to calculate the categorical coefficient for less than 50% of the groups. From this table, the sinusoid coefficients are the coefficient of the sinusoid, and the categorical is the standard added value for the measured SWF values in that month (subtract it for the correction). These values are not significantly correlated for the months of November and December. All the other found values are significantly correlated. The coefficient for the sinusoid is the amplitude which is about the same as in Figure 4-5.

Coefficient SWF	*Sinusoid Day*	*Sinusoid week*	*Sinusoid month*	*Categorical interval month*							
				May.	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.
Coefficient	0.05284	0.05167	0.05107	-.0277	-.0581	-.0554	-.0747	-.0720	-0.052	-0.006	0.009
P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.268	0.101

Table 4-5 Liner coefficient and categorical coefficient date interval

The correction is necessary since the correlation tests for a linear relation; this also changes the Pearson Correlation from a number that indicates a linear relation to a sinusoid relation. The problem is that the phase shift is now based on the calculation found in Figure 4-5, and might fit better with another phase shift after correcting the SWF with e.g. temperature. Since the sinusoid is currently used in the model as a function of date and fits significantly with the current phase shift calculation, it should be included in the model. In the end model, a new phase shift can be found in the appendix.

4.4.2. Conclusion seasonal variation

The results of the test show that there is a significant correlation between the sinusoid of date number and the SWF coefficient. The proof for a linear relationship between these variables was insignificant to determine this relationship, which was expectable for a total time interval of a year. The seasonal variation as time interval can be expressed through month number, week number and day number. From the results of Table 4-4 can be concluded that every interval is significantly correlated. Day number would have the most exact expression for the seasonal variation but is also the most influenceable by outliers during the determination step of finding the coefficients for the sinusoid. However, if the month number gets chosen as an expression for this interval, the correction can be very off if the measuring date is on the first or last day of the month. Since the month coefficient can be influenced by outliers, we choose for an exact correction to keep the confidence interval small.

In Table 4-4 the relation between the time interval and the other input variables is shown. From this table can conclude that the amount of rain and the temperature is correlated with the day number. Only the variable days before the measuring day where the amount of rain was more than 25mm is not significantly linear correlated. Since there is a significant correlation between most these variables, it is extra important to determine the coefficients of all the input variables in one model.

5. Formulating a model

The goal of this chapter of the project is to determine possible models and choose the best one. The model will exist out of the correction formula and the restrictions for measuring. In the last chapter, the correlation between the individual variables and the SWF value was determined. In this chapter, the correlation between the multiple variables and the SWF value will be determined. This multiple regression analysis determines the influence of the variables and which variables should be included in the correction formula. The models will be built and evaluated based on four aspects.

1. The relevancy of an additional predictor.

The relevancy of a model for linear regression, as in Chapter 2.2, was based on the t-value. The t-value for multilinear regression should not be used since many predictors are fitted. With many predictors, there will always be some that have a small p-value even though they might not be statistically significant. Therefore, the F-statistic is used, this is based on the amount of data and variables and used to calculate a P-value. This determines if the addition of the variable in the model is of significant influence. [19]

2. The multicollinearity of the predictors.

An import aspect of the addition of a variable is multicollinearity**.** Multicollinearity is an expression for the correlation between the predictors. If this correlation is high, it is hard to determine the best predictor and its coefficient. Therefore in this step, it is important to take into account what the sign of a variable was in Chapter 2.2, and that it should be the same in the total model. This does not mean the influence of an individual variable could not be negative if found to be positive now. The variable can be compensated by the influences of correlated effects. If this would be an option, the model might have a perfect R-squared(accuracy) with this data. However, when it is used with data where the input variables (predictors) are not as highly correlated, the accuracy of the model can differ a lot. Therefore, multicollinearity lower than six is preferred. The max VIF of a model will be used for assessing the model. [20]

3. The additional accuracy of the model.

The R-squared can be used to analyse the accuracy of a multiple linear regression model. However, in multiple linear regression analysis, the addition of an extra variable always leads to an increase in R-square. This happens because the coefficient of the variables can be optimally chosen to get the best results. This can be logically concluded if for example, a model with currently one variable is chosen. The addition of another variable will at least have the same accuracy. In the worst case the variable can have a coefficient of 0, thus not increasing the R-squared, but never decreasing it.
Therefore, stepwise regression and best subsets regression can be used to make a trade-off between precision and bias. The overall model will be assessed on the standard deviation.

4. The complexity of the model.

The last character on which the complete model is assessed is its complexity. The input variables needed for the calculation can be measured or expressed in multiple ways. The occurrence of rain in the last week can, for example, be expressed in it did or did not happen, but also in the amount of rain. Some expressions are easier to measure than others. This is also an important aspect since the additional accuracy of the model by an input variable can be significant but make it practically unusable. Therefore, the whole model and its restrictions will also be assessed on its practicality.
The model will be graded between 1 (not complex) and 5 (very complex), with an explanation.

5.1.2. The influence of rain

In the next part of the project, seven models are determined that can correct the measured SWF value. The last models use rain as an input variable for the correction. In Chapter 4.3 it has already been discussed that rain can be expressed and measured in many ways. To use rain as input variable, the measurement of rain is limited to keep the model simple. The second restriction is the data which was available for this research. The data of the rain stations only measured the amount of rain in a period of 24 hours.

Chapter 4.3 concluded that there was no linear relation between the amount of rain in some of the previous days and the measured SWF. However, based on literature and experience of RWS, there is a relation between the cleanness of the road and the SWF suggested. Therefore, a minimum amount of rain in a certain time period can clean the roads which influence the SWF. The influence of a dirty road has been considered by restricting measurements if no more than 1 mm rain occurred in a 24-hour interval in the last 14 days. Rain always makes the model complex since it is hard to measure. Because most roads are not close enough to a “neerslagstation”, the rain must be measured by the companies in advance. Measuring the amount of rain over the span of a road takes a lot of additional work.

5.2. Making a model

The models are made with the use of Minitab, and mainly the function of multiple regression analysis. To start this, analysis variables are selected for input. These variables are divided into two groups: continuous variables and categorical variables. Continuous variables are the variables used for direct input of the correction model and have a continuous scale. Categorial variables are used for the different groups in the model or analysis, e.g. the measurements places. The use of some categorical variables, like measurement places, is necessary since the actual SWF differs per group. An example of these results from a multiple regression analysis is given in the appendix.

The multiple regression analysis finds the optimum coefficients for all the variables to explain as many measurements as possible. Therefore, it is important that we first identified which variables are relevant to avoid overfitting.

For the formulation of the models, the coefficients of the multiple regression analysis are used. These coefficients give the correlation between each variable and the measured SWF. For the evaluation of the results, the R-squared, maximum p-value, maximum VIF and standard deviation are used from this analysis. The complexity of the models is determined based on the variables which are used.

5.3. Evaluating a model

In the introduction, there is already an explanation provided on which aspects a model is assessed. In Chapter 2, the influence of independent variables is determined. In the next part, the results of the multiple regression analysis are given by the formulation of multiple models. An explanation with advantages and disadvantages for each different model is given. The choice of variables in each model is based on the findings of Chapter 2 and Chapter 4.

5.3.1. Model one (Twater Troad)

$$SWF_c = SWF + 0.0042 * (T_{water} - 20) + 0.001 * (T_{road} - 20) - 0.0169 \sin(\frac{2\pi}{365}(x - 7))$$

Temp water	Temp Road	Temp Tyre	Temp Air	Sinusoid day	Days no rain
-0.0042	-0.001			A=.0169 β=7	

Table 5-1 Variables included model one

The first model that will be made with the current data contains the predictors which Rijkswaterstaat currently uses. For there calculation, they use the temperature of the road, the water and the sinusoid for the seasonal variation. The coefficients for their corrections are different, which can be explained due to the different data groups used for the analysis.

R-square	Max P-Value predictor	Max VIF	Complexity	Standard dev
92.73%	0.000	3.01	1	0.0168

Table 5-2 Summary results model one

Overall, this model seems very good, with an explained variation of 92.73% using this correction. Only 3 predictors are used which use data that is measured during measurements. Therefore, the complexity of gathering this data is low and therefore, the model has relatively low complexity. The multicollinearity is below 6, and the relevancy of every predictor is significantly proven by the F-statistic and thus P-value.

5.3.2. Model two (Twater)

$$SWF_c = SWF + 0.0058 * (T_{water} - 20) - 0.0154 \sin(\frac{2\pi}{365}(x + 4))$$

Temp water	Temp Road	Temp Tyre	Temp Air	Sinusoid day	Days no rain
-0.0058				A=.0154 β=-4	

Table 5-3 Variables included model two

The second model is an adaptation to the original model. Here the road temperature is not considered since in the analysis of the individual variables already was discovered that they are highly correlated. Here the conclusion is made to take only 1 or 2 temperature variables.

R-Square	Max P-Value predictor	Max VIF	Complexity	Standard dev
92.05%	0.000	1.4	1	0.0176

Table 5-4 Summary results model two

This model has good accuracy for only using one variable. Since only one temperature is used as input, the highest VIF has a low value of only 1,4. However, the accuracy and standard deviation are not as good as the original model based on this data set. The choice for this model, over the original model, could be based on the trade-off between the accuracy or the relevance of the additional variable. For now, the conclusion is that this model is a good alternative.

5.3.3. Model three (Troad)

$$SWF_c = SWF + 0.0023 * (T_{road} - 20) + 0 * (T_{tyre} - 20) - 0.0311 \sin(\frac{2\pi}{365}(x + 30))$$

Temp water	Temp Road	Temp Tyre	Temp Air	Sinusoid day	Days no rain
	-0.0023	insignificant		A=.0311 β=-30	

Table 5-5 Variables included model three

This model uses only the road temperature and looks if tyre temperature is a good addition as some literature suggested that this might have the most significant relationship. However, with this data, we know that the tyre temperature is measured differently. Tyre temperature does not have a significant influence in all our multi-linear regression models.

R-square	Max P-Value predictor	Max VIF	Complexity	Standard dev
91.61%	0.332 for Tyre (others: 0.000)	1.11	1	0.01913

Table 5-6 Summary results model three

This model uses only road surface temperature since the regression analysis shows that the addition of tyre temperature is not relevant. The model has a very low MIV but a high standard deviation. The sinusoid used in this model also has a high value for its amplitude which means the influence of the day number is higher than usual. The problem of a higher amplitude is that it only is an expectancy of the circumstances. Thus, indicating the influence, we expect for that day, therefore does not take deviant circumstances into account. Therefore, this model is not highly recommendable.

5.3.4. Model four (Tair)

$$SWF_c = SWF + 0.0031 * (T_{air} - 20) - 0.0310 \sin(\frac{2\pi}{365}(x + 34))$$

Temp water	Temp Road	Temp Tyre	Temp Air	Sinusoid day	Days no rain
			-0.0031	A=.0310 β=-34	

Table 5-7 Variables included model four

The last model with only temperature and seasonal variation included is that with only Air temperature. The temperature of air is one of the most variating temperatures and is a direct influence or directly influencing the circumstances. However, air temperature is also very unstable during the calculations and has a lot of variation. Therefore, in Chapter 4.2 was concluded that water and road might be better. This model is made to test if the same can be concluded with this dataset.

R-square	Max P-Value predictor	Max VIF	Complexity	Standard dev
89.48%	0.00	1.13	1	0.02025

Table 5-8 Summary results model four

The results of this model are good but not as good ad the previous results. The model does have low complexity, but this has a negative influence on the standard deviation, which is high. Therefore, the dataset also concludes that the temperature of air is not the best input variable.

5.3.5. Model five (Twater Troad Rain)

$$SWF_c = SWF + 0.0040 * (T_w - 20) + 0.0012 * (T_r - 20) - 0.0179 \sin\left(\frac{2\pi}{365}(x - 3)\right) + 0.0072 * B$$

Temp water	Temp Road	Temp Tyre	Temp Air	Sinusoid day	3 Days 1< rain
-0.0040	-0.0012			A=.0179 β=3	-.00716

Table 5-9 Variables included model five

This model is the original model with the addition of rain. The addition of rain as a dummy variable in the model leads to a decrease in the coefficient of water temperature. The change in variables also sets the starting date of the sinusoid a little bit further away but not significantly. The increase of the sinusoid suggests that rain had a reversed or unstable influence as a function of time.

R-square	Max P-Value predictor	Max VIF	Complexity	Standard dev
92.95%	0.002 (3 Days 1mm<rain)	3.16	4	0.0166

Table 5-10 Summary results model five

The results show the higher R-squared with a low standard deviation. This model is comparable with the original model, not only the results but also the variables that are used. There is a small increase in the R-squared and decrease in the standard deviation, however, this model is way more complex. It has a high complexity since it is hard to measure rain accurately. Since a lot of measurement places, and even more roads, are not near measurement places. Thus, the rain must be measured in advance of measuring the SWF of a road.

5.3.6. Model six (Twater Rain)

$$SWF_c = SWF + 0.0057 * (T_w - 20) - 0.0164 \sin\left(\frac{2\pi}{365}(x+8)\right) + 0.0050 * B$$

Temp water	Temp Road	Temp Tyre	Temp Air	Sinusoid day	3 Days 1< rain
-0.0057				A=.0164 β=-8	-0.0050

Table 5-11 Variables included model six

This model uses only water temperature instead of water and road as in the previous model. The exclusion of road temperature leads to an increase in the influence of water temperature. Another big difference is the influence of rain that significantly decreased.

R-Square	Max P-Value predictor	Max VIF	Complexity	Standard dev
92.16%	0.035 (3 Days 1mm<rain)	1.44	4	0.0175

Table 5-12 Summary results model six

The results of this model are again like the last model. This agrees with some literature which states that only the influence of one temperature is needed for a correction. However, the model also shows a large increase in the P-value of the influence of rain. Therefore, it is less like that rain is a significant influence in this model. The complexity is a bit lower since only one variable for temperature must be taken. However, the temperature of water, road and air are measured using the same method; therefore, it does not matter a lot.

5.3.7. Model seven (Twater Troad Rain)

$$SWF_c = SWF + 0.0042 * (T_w - 20) + 0.0011 * (T_r - 20) - 0.0179 \sin\left(\frac{2\pi}{365}(x-1)\right) + 0.0061 * B$$

Temp water	Temp Road	Temp Tyre	Temp Air	Sinusoid day	5 Days 2< rain
-0.0042	-0.0011			A=.0179 β=1	-.00608

Table 5-13 Variables included model seven

The last model shown in this report is one where the influence of rain in the model is considered for 5 days instead of 3 days. This again is a dummy variable that is 1 if it has not rained in the last five days more than 2 mm within 24 hours. The model again has similar results for the coefficients of variables.

R-Square	Max P-Value predictor	Max VIF	Complexity	Standard dev
92.88%	0.010 (5 Days 2mm<rain)	3.02	5	0.0167

Table 5-14 Summary results model seven

The results of this model are like model five. However, the complexity is higher since now the data of not 3 but 5 days or rain must be measured. Therefore, this model has almost the same accuracy, and standard deviation should not be chosen over the previous model.

5.4. Choosing a model

In the last part, seven models were formulated. The goal of this research is to find the best model for the correction. The best, in this case, means a model which has the best prediction of the "real" SWF. Therefore, the most important aspects of the model are the R-squared and the standard deviation. However, the model should also be practical to use and thus use practical variables. In this part, the best model is chosen based on these aspects. Important note for this decision is that the model is not finished when only the formula is chosen, in addition, restriction and requirements for the model can be set.

5.4.1. The options

Of the seven models formulated in part 5.2, models **one**, **two**, **five** and **six** will be compared.

Model **three** is excluded since it has lower results than model two. Both do use only one variable since tyre temperature was excluded, but the temperature of water seems to be a more valuable predictor. The influence of the seasonal variation is stronger than most models; this is only a predictor of the circumstances on that day, which we want to keep small.

Model **four** is excluded because it also has some lower results. In Chapter 4.2 we concluded that the temperature of water and surface was more relevant as a predictor. Here again, the influence of the seasonal variation is stronger than most models; this is only a predictor of the circumstances on that day, which we want to keep small.

Model **seven** has some very promising results. This model uses rain as a direct input for the model. To properly use this model, the amount of rain in the five days before measuring should be measured. The use of rain over a period of 5 days makes it more complex than over a period of three days. Since model five also has better results than this model, we exclude model seven.

aspect	*Model 1*	*Model 2*	*Model 5*	*Model 6*
R-square	92.73%	92.05%	92.95%	92.16%
Standard deviation	0.0168	0.0176	0.0166	0.0175
Added R-squared by rain	--	--	0.22	0.11
Complexity	1	1	4	4
Highest VIF	3.01	1.40	3.16	1.44
Respectively multicollinearity	3rd	1st	4th	2nd
Number of variables	3	2	4	3
Most difficult variable to measure	Road temperature	Water temperature	Rain in the past 3 days	Rain in the past 3 days

Table 5-15 Summary of models

Table 5-15 Summary of models shows that **model 5** has the **best results** in accuracy since it has the highest R-squared and lowest standard deviation. However, the model also is one of the most complex ones with the highest multicollinearity. The complexity makes it harder to use since more measurements must be done to calculate the results. The multicollinearity can lead to large deviation in an unusual situation, and this is a situation that did not occur when making the model. An example of such a situation can be a high temperature for water but low for road surface.

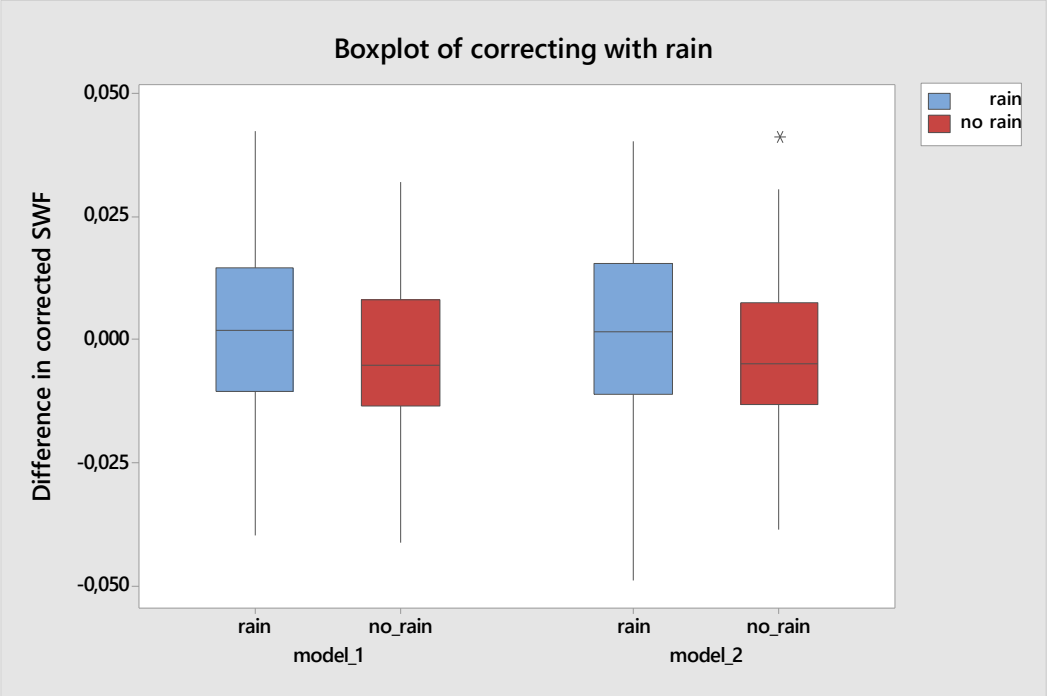

Figure 5-1 Boxplot of Corrected SWF values by model 1 and 2 vs the occurrence of rain in the past 3 days

In Figure 5-1 the effect of excluding rain is given after correcting the measured values with model 1 and 2. In these boxplots, the difference between the corrected SWF value and the expected SWF value is analysed. This shows that a remaining difference in SWF can be explained by the occurrence of rain in the past three days. We conclude that on average, the SWF values after a short period of drought are corrected lower than they should have been. However, on average, the difference in this correction is very small for both these models.

The difference between the R-squared of model 1 and model 5 is 0.22. This difference is the result of the addition of rain in the model as an input variable. We do acknowledge that rain has a significant influence. Nevertheless, the additional work to correct these measurements negatively influences the complexity. With the current results of using rain as an input variable, we would **not** recommend it. The trade-off between the work of measuring the amount of rain over such a large area and a small increase in accuracy does not seem to be worth it.

Now we are left with **model 1** and **model 2,** only a few small differences can be noticed. The accuracy (R-square) has a small difference between the models. The difference in standard deviation is a bit higher, respectively. However, model 2 does score a lot better in multicollinearity, where it is first. Model 2 limits itself only to the use of water temperature; model 1 additionally uses road surface. This addition leads to an increase in the multicollinearity. What we did not expect was a decrease in the influence of seasonal variation if the temperature was only limited to water. This suggests that if road surface temperature is not used, the influence the predictor for seasonal variation decreases.

5.5. The recommended model

Based on the conclusion and results in the previous chapters, a new model can be selected. Both are good options and easy to use. Model 1 is already currently used (with a slightly altered formula since a larger dataset was used.) Both have an acceptable value for the multicollinearity. We would **recommend model 2**. The results of Chapter 4.2 and the literature results that only one predictor for temperature is necessary. The decrease in the coefficient of the sinusoid also seems like a positive

effect on the correction formula. A higher coefficient might be the results of overfitting, where the addition of every variable would always lead to an increase. It has a low VIF, one of the highest R-square, does not use complex variables and has a small coefficient for the sinusoid. A disadvantage of using this model can be an exceptional situation where the temperature of the water is very different than the temperature of the road surface. This can be a problem when road temperature has a stronger influence on the SWF than currently expected.

The model which uses *water* and *road* temperature is a safe option. Two of the variables in this model are highly correlated, road surface temperature and water temperature. This model, of course, has a higher accuracy with the current dataset. However, it is hard to determine if the higher accuracy is a result of overfitting or is a good contributor in predicting the SWF.

Currently, a model with water temperature and road surface temperature is being used. We did not find the same results to choose a similar model. We agree that the differences are very small currently and know that the original model is already accepted as a reliable correction method. Therefore, we conclude that **model 2** would have **our preferences**, yet can we agree with **Rijkswaterstaat** if they want to keep using the **original model**. The advantage of this model can be that it was based on a previous data set with more data that was not limited to measurement places near a rain station.

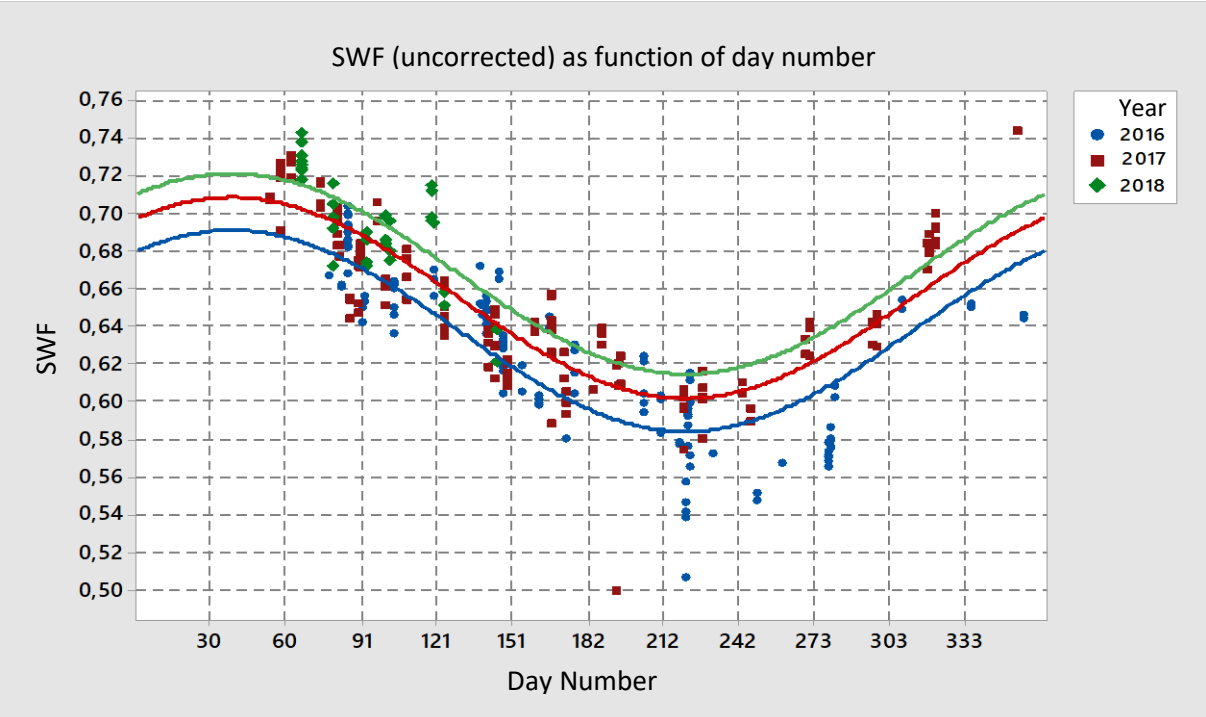

Figure 5-2 SWF measured values as function of days

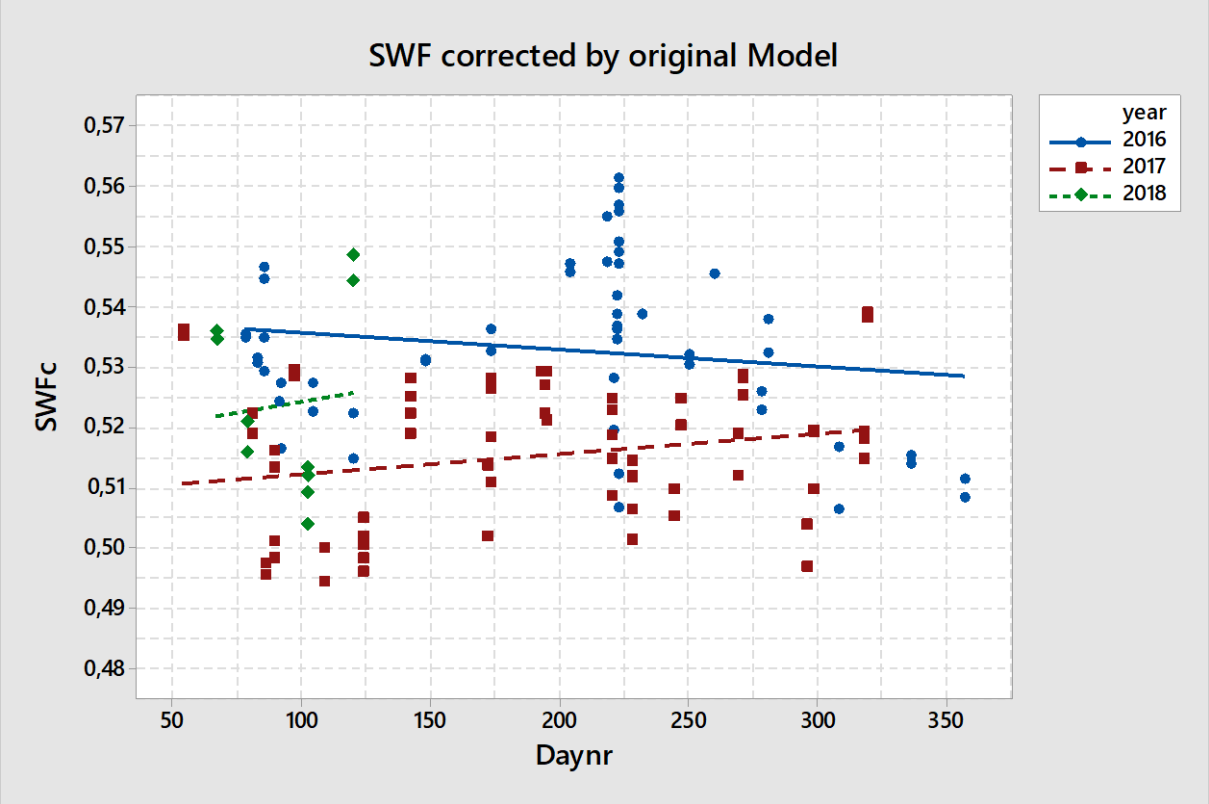

Figure 5-4 SWFc example original model

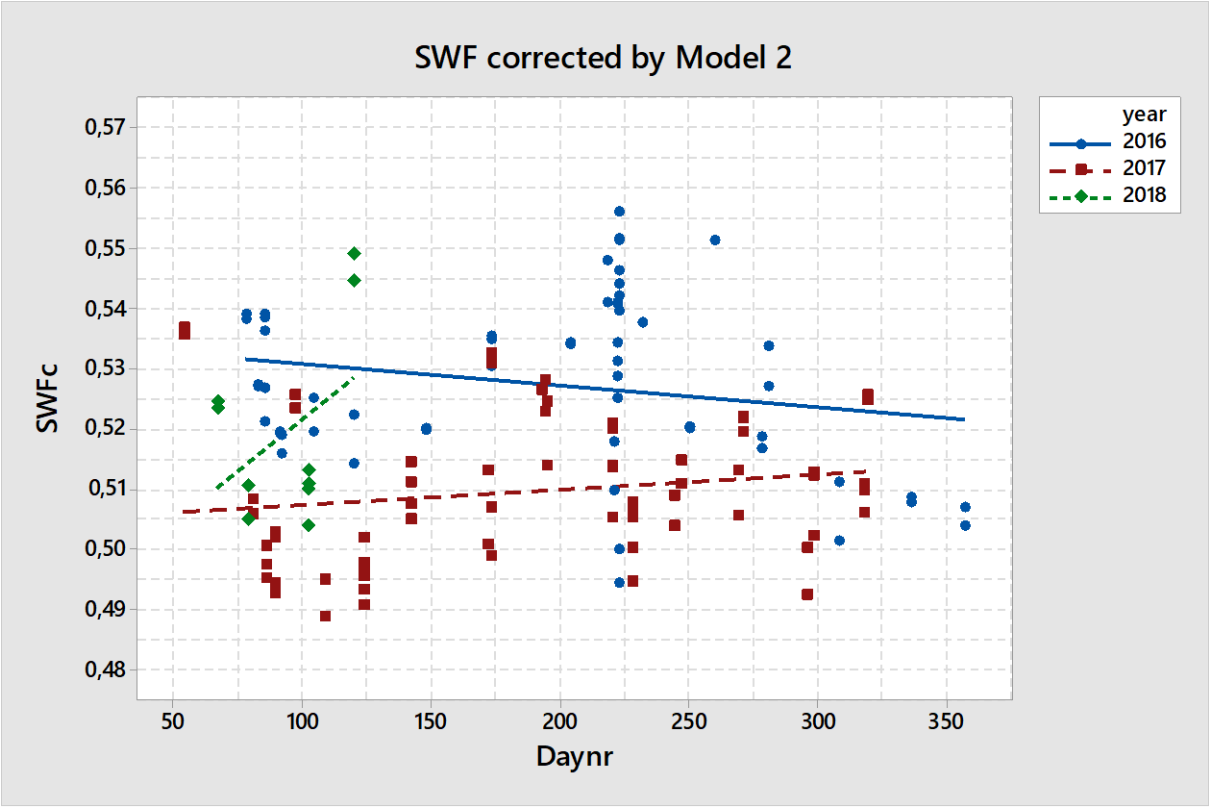

Figure 5-3 SWFc example Model 2

.

In Figure 5-4 and Figure 5-3 a scatterplot is made to show how the SWF values are corrected for each day. The plotted lines are linear functions for each year and should represent the expected SWF value for that year. For these figures, the same data is used as in the introduction for Figure 1-3. We already mentioned an unexpected increase in SWF for each year.

The correction model shows that this increase is explainable by the influences we determined. We see that after correcting the measurements, they are lower than in 2016. The expected value of 2018 is based on the correction of 5 measurements which could explain the high slope.

For 2016 and 2017 we do see a small slope, but it overall looks stable. If this correction formula were used on new data, we would expect a small decrease in slope over the year. This represents the expected SWF value decreasing due to traffic.

The correction of both models is similar if we look at the overall results. We do notice that the corrections by model 2 are on average, resulting in lower corrected values than the original model. According to assumption 0, the corrected value should be close to the actual value. Both models use different coefficients for the variables, and the original model is determined with the use of another data group. This could explain the difference between the corrected values.

6. Increasing reliability of the model

In this part, the goal is to determine which restrictions are required to make the model as reliable as possible. This chapter is divided into two parts.

The first part analyzes the use of the drought restriction. This was a request of Rijkswaterstaat; they wanted to know whether their current restriction is suitable. This restriction requires minimally 1 mm of rain within one day in the last 14 days. This restriction is set to ensure that the road surface is not soiled. This restriction seems like the use of a dummy variable we already included in some of the models.

The second part is further recommendations for research to keep the model relevant or help with formulating a better model.

6.1. Restriction of drought

In the current dataset, the five measurement places have a maximum period drought (less than 1 mm rain each day) of 12 days. Therefore, all the measurements satisfy the current requirement. Table 1-1 gives the maximum amount of days before a certain amount of rain has fallen. We can see that every measurement had more than 2 mm of rain on a day within 21 days before measuring.

Amount of rain on one day	*0,1*	*0,5*	*1*	*1,5*	*2*	*2,5*	*3*
Variable Boolean	D01	D5	D10	D15	D20	D25	D30
Maximum amount of days	10	12	12	12	21	24	24

Table 6-1 Maximum amount of days before the minimum amount of rain in one day occurred

Since the dataset for this research only includes measurements that met this restriction, no analysis to extending the number of days with this amount of rain can be done. We can analyze the result of shortening the period with the current amount or extending with a higher amount.

6.1.1. reducing the day limit of drought

In this part, we investigate if shortening the period or changing the amount of rain needed within one day to consider a road to dry, significantly changes the results of the correction formula. To do this, the SWF values are corrected with the formula of the **original model**, **model 1** and **model 2**. Then the difference between the **measured SWF** and the **expected SWF** is calculated and defined as **nSWFc**. These SWF values will be used to analyze the regression between the drought and the remaining difference after correcting.

For the regression, they are divided into the following groups with a Boolean variable D10. This variable can have a value for D up to 12, where D is the number of days before a minimum of 1.0 mm of rain has fallen within one day. This gives two different groups, one that does and one that does not satisfy this Boolean. An example of these results are given in Table 6-2, here the difference of the standardized SWF is calculated for the first seven days where the amount of rain should be at least 1 mm on a day. Here standardized means that it recalculated to make it comparable with respect to the measurement place, temperatures, date and expected actual SWF value.

D10 (1 mm of rain)	**nSWFc_original**	**nSWFc_model 1**	**nSWFc_model 2**
Day 1 Pearson value	-0.222	-0.172	-0.223
Day 1 P-value	0.000	0.002	0.000

Day 2 Pearson value	-0.202	-0.185	-0.156
Day 2 P-value	0.000	0.001	0.004
Day 3 Pearson value	-0.169	-0.146	-0.119
Day 3 P-value	0.002	0.007	0.030
Day 4 Pearson value	-0.188	-0.167	-0.113
Day 4 P-value	0.001	0.002	0.039
Day 5 Pearson value	-0.185	-0.164	-0.104
Day 5 P-value	0.001	0.003	0.057
Day 6 Pearson value	-0.174	-0.144	-0.103
Day 6 P-value	0.001	0.008	0.059
Day 7 Pearson value	-0.076	-0.062	-0.046
Day 7 P-value	0.162	0.256	0.399

Table 6-2 D1.0 groups of rain occurrence

The same calculations are done with different amounts of rain for D0,1, D0,5, D1,5, D2,0 D,2,5 and D3,0. For all the values and models no significant effect between amount of rain and the measured SWF was found. Here again, it should be noted that the amount of measurements with a period of longer than five days is already lower than 20% for every rain amount. The Pearson values are not very high; thus, we conclude influence that the influence is **significant** but **relatively low** for the standardized SWF values.

In Table 6-2 rain does not seem a significant factor for model 2 at six days. An explanation for this can be that when the correction is performed with model 2, which does not use road temperature, that the groups do not seem significantly different. The multicollinearity between the variables of model 2 and rain can lead to a decrease in a significant difference between these two groups.

6.1.2. Results of reducing the limit

For the original correction model and model 1, a limit of 6 days of drought gives a significant influence on the results. Both model 1 as the original use the same variables but they have different coefficients. Therefore, a new model is created that uses the seasonal variation, temperature of water and road surface. However, now only the data is used where there was a minimum amount of 1 mm rain in the six days before measuring.

R-Square	Max P-Value predictor	Max VIF	Complexity	Standard dev
92.90%	0.000	3.11	4	0.0170

The results of this model are **slightly better** than model 1 since the **R-squared** is higher; however, the **standard deviation** is also **higher**. The model is **more complex** than usual since the reduction shortens the period in which measurements can be performed.
Based on this small increase in R-squared, the best conclusion would be not to **shorten** the restriction of drought since it reduces the dataset by 25% and only has a small influence on the results. Therefore, it limits the measuring time for a small increase in the R-square. In addition, the standard deviation is higher due to the lower amount of measurements.

R-Square	Max P-Value predictor	Max VIF	Complexity	Standard dev
91.79%	0.000	1.46	4	0.0179

Model 2 has also been recalculated with the excluded values which did not fulfil the requirement of rain within a period of 5 days instead of 6 days. For this model temperature of the water and the influence of day number is used. The result in R-squared is slightly better than model 2, but again the standard deviation is worse. Therefore, the conclusion is again that the restriction of drought should not shorten when using model 2 since it makes the requirements for measuring harder and thus gives a more complex model while the R-squared is not significantly increased.

6.1.3. Increasing the rain amount

D10 (1 mm of rain)	nSWFc_original	nSWFc_model 1	nSWFc_model 2
D15 15 Pearson value	*	*	*
D15 15 P-value	*	*	*
D20 15 Pearson value	0.04	0.049	-0.006
D20 15 P-value	0.461	0.37	0.917
D25 15 Pearson value	0.098	0.12	0.072
D25 15 P-value	0.073	0.028	0.186
D30 15 Pearson value	0.098	0.12	0.072
D30 15 P-value	0.073	0.028	0.186
D15 16 Pearson value	*	*	*
D15 16 P-value	*	*	*
D20 16 Pearson value	0.076	0.078	0.039
D20 16 P-value	0.164	0.153	0.474
D25 16 Pearson value	0.13	0.15	0.118
D25 16 P-value	0.017	0.006	0.031
D30 16 Pearson value	0.13	0.15	0.118
D30 16 P-value	0.017	0.006	0.031
D15 17 Pearson value	*	*	*
D15 17 P-value	*	*	*
D20 17 Pearson value	0.076	0.078	0.039
D20 17 P-value	0.164	0.153	0.474
D25 17 Pearson value	0.13	0.15	0.118
D25 17 P-value	0.017	0.006	0.031
D30 17 Pearson value	0.13	0.15	0.118
D30 17 P-value	0.017	0.006	0.031
D15 18 Pearson value	*	*	*
D15 18 P-value	*	*	*
D20 18 Pearson value	0.076	0.078	0.039
D20 18 P-value	0.164	0.153	0.474
D25 18 Pearson value	0.086	0.09	0.066
D25 18 P-value	0.114	0.1	0.226

D30 18 Pearson value	0.086	0.09	0.066
D30 18 P-value	0.114	0.1	0.226
D15 19 Pearson value	*	*	*
D15 19 P-value	*	*	*
D20 19 Pearson value	0.076	0.078	0.039
D20 19 P-value	0.164	0.153	0.474
D25 19 Pearson value	0.086	0.09	0.066
D25 19 P-value	0.114	0.1	0.226
D30 19 Pearson value	0.086	0.09	0.066
D30 19 P-value	0.114	0.1	0.226
D15 20 Pearson value	*	*	*
D15 20 P-value	*	*	*
D20 20 Pearson value	0.076	0.078	0.039
D20 20 P-value	0.164	0.153	0.474
D25 20 Pearson value	0.086	0.09	0.066
D25 20 P-value	0.114	0.1	0.226
D30 20 Pearson value	0.086	0.09	0.066
D30 20 P-value	0.114	0.1	0.226
D15 21 Pearson value	*	*	*
D15 21 P-value	*	*	*
D20 21 Pearson value	0.076	0.078	0.039
D20 21 P-value	0.164	0.153	0.474
D25 21 Pearson value	0.086	0.09	0.066
D25 21 P-value	0.114	0.1	0.226
D30 21 Pearson value	0.086	0.09	0.066
D30 21 P-value	0.114	0.1	0.226

Table 6-3 Significance of adding drought as restriction to the models

In Table 6-3 is calculated if there is a significant difference between normalized SWF with a longer period and higher amount of rain. These calculations cannot be done for D01, D05 and D10 since the current restriction is a period of 14 days for D10. Since a restriction in the model makes it more difficult to use an additional research is done to see if the period in the limit can be set further away. Again, the problem is that the current restriction does not allow measurements if the amount of rain has been lower than 1 mm for every day in the last 14 days (D10 14days). Since all the data fulfils this requirement, the only research that can be done to extend the limit, is with another amount.

Table 6-3 Significance of adding drought as restriction to the models shows * for all the D15 values; this means that there is only one group. Thus, all the requirement that fulfil D10 also fulfil D15 for the same time period. And thus, no calculation between two groups can be done since they all fulfil the requirement.

All the P-values are relatively low. To extend the drought restriction, we want high results for the p-value since that proves that there is not a significant difference between the two groups. Note that there are relatively few measurements that do not fulfil this restriction. Therefore, it is hard to prove a significant difference.

The Pearson value is low; this indicates the strength of the relation. It seems logical that there is a weak relation between these values since we already know that at least 1 mm rain has occurred in the previous 12 days. Therefore, these measurements can be considered less dry by this occurrence.

The restriction of drought could be increased to have a longer-term if the drought gap would be set on 2 mm instead of 1 mm. However, we would recommend also leaving this unchanged since there is a relatively low amount of data available with long term drought before measuring. We would recommend keeping the old restriction. This can again be revises when research has been done with measurements that do not fulfil the requirement of minimal 1 mm of rain on a day within 2 weeks.

6.2. Recommended research

For this research, we concluded with a recommended model. This model uses the temperature of water and the seasonal variation. During the calculations, some conclusions were already made because of the lack or incorrect data. In this part, we recommend some follow-up research that could be relevant.

6.2.1. Temperature water and road surface

In the last part of the remaining choice was between two models that both were similar, except for the inclusion of road surface temperature. In the start of this research, it was already concluded that the temperatures were highly correlated. Therefore, it is hard to determine the exact influence of each of these temperatures. Since the original model uses both temperatures, we would suggest determining the individual influence of these temperatures.

To perform such research, we would recommend measuring in the summer when the temperatures are hot. Instead of using water from nearby sources during these measurements, we would suggest using cooled tapped water. Then measurements are performed where the two temperatures are not correlated. This can confirm if the original model or our model should be used.

6.2.2. Tyre temperature

In an early stage, we discovered that there was an unexplained difference between the temperature of the tyre in 2014 and 2016. The relation between tyre temperature and other variables, including SWF and the sinusoid of date, was different in each year. After consulting with RWS, we determined that the measurements were performed by two different care types. This suggested that the old car used a different kind of temperature sensor.

The literature does suggest that tyre temperature could have an influence on the skid resistance. However, even with the corrupted data, we determine a high correlation between the temperatures. Therefore, we would suggest that if additional research would be performed, the measurement method of tyre temperature should be verified according to the established protocol.

6.2.3. Extend the restriction of drought

One of the questions of Rijkswaterstaat was to determine if the restriction of drought is too short. For this question, we had to see if there was a difference between measurements that did not meet this requirement at a certain point. In the report, we did analyze the difference between

measurements that did not fulfill the earlier than the requirement of 14 days. We also analyzed the difference between measurements that did not fulfil the drought requirement but with a higher minimum for rain.

However, since all the measurements did fulfil the requirement, it is not possible to analyze a difference between these groups. Therefore, we can not determine if the restriction could be extended with this data. If this restriction is a problem for the measuring companies, we would suggest performing measurements when this restriction is not met. This could help to determine if there is a significant difference after a period of drought in the skid resistance.

7. Conclusion

In this report, the main research was focused on, identifying influences of weather variables and formulating a new model with this data. The goal was to formulate models with high accuracy but low complexity. In this report, research questions were formulated to solve this goal. In this chapter, we summarize the answers to these questions.

7.1. How to determine the influence of each independent variable

This question helps to determine which variables have an influence on the SWF and should be included in the formulation of the models. To solve this, multiple questions were formulated; these questions are answered in the literature research and applied in Chapter 4.

We started with *identifying possible short-term influences on the measured SWF*. This part has been done in literature research. We concluded that temperature, seasonal variation, rain and drought could have an influence on the measured SWF.

Next, we *specified the scale and measurability of the independent variables*. Here again, the literature study suggested what had an influence of the skid resistance. However, it still was important to use the right expression for the variables. We concluded that temperature could be expressed in the measured Celsius. Seasonal variation indicated by the day number in the function of a sinusoid. The rain as a dummy variable, expressed in a minimum amount of rain had occurred in the past X days.

At the end of the literature review, we discussed which variables could be used to *evaluate the correlation per variable and the SWF*. Here we concluded that the Pearson correlation coefficient helps to determine the strength of the relation. The P-value was used to determine if a correlation can be significantly proven and the R-squared for the accuracy of using the predictor.

7.2. Which variables should be chosen for a properly working model?

The goal of this research question is to determine whether a variable should be used in the model to help with the accuracy, without making it too complex.

In Chapter 4, the correlation between each variable and with the SWF has been determined. The *correlation between each variable and the outcome* determined the accuracy each variable could add to the model.

For all these variables, we determined that they had a significant influence on the measured SWF. However, we also concluded that the number of variables should be limited to prevent overfitting.

In the meetings with Rijkswaterstaat, was discussed that rain as input variable was a very complex variable. Since rain should be measured in advance of measurement and only had a small influence on the measured SWF, *the complexity did not outweigh the significance of the outcome*. The other variables were not determined as a complex variable by Rijkswaterstaat. Their values were automatically determined during the measurements. We did conclude that tyre temperature should not be concluded since the measurement method was different per car type.

For the use of temperature, we concluded that only one or two of the temperatures should be used as a direct input variable. All the temperatures were highly correlated; this would suggest overfitting, seen by high VIF values if multiple would be used.

7.3. How to choose the best model for correcting the SWF

In this part of the project, multiple correction formulas were formulated. Every formula was determined with the current dataset.

At the beginning of Chapter 5, we introduced the indicators to *evaluate the quality of a model*. These are the relevancy of the predictors used in the model, the multicollinearity within the model, the accuracy of the variables in the model and the complexity of the model. In the literature search, these indicators were found and explained.

In the next part, the models were formulated. For this formulation, multiple linear regression was performed with the variables. Each variable was added stepwise, and at the end, the coefficients were determined. All the models were then compared to their *advantages and disadvantages*.

The models that used only road surface and air temperature are excluded first. They have lower accuracy than the models that included the temperature of the water. In addition, they have a higher coefficient for the sinusoid; this is only an indicator of the expected seasonal variation of that day. Therefore, these have a weak spot for unexpected seasonal influences. Next, we concluded that the models with rain as the input variable are more complex. They did have a higher accuracy; however, this could also be because of overfitting. The added accuracy of these did not outweigh the complexity.

We remain with model 1 and model 2. Model 1 uses the same variables as the current model of Rijskwaterstaat. The coefficients in this model are different since we used a different dataset. Our recommended model would be in model 2. The *advantages* of model 2 over model 1 were as literature, and our analysis suggested that only one temperature was necessary. A lower multicollinearity, the lower influence of seasonal variation expressed in the sinusoid and the smaller chance of overfitting. We do acknowledge that model 1 has the following advantages over model 2; A higher accuracy, a lower confidence interval, better suitability when water temperature and road surface temperature are not correlated. In addition, it is already used in practice and proves to be sufficient.

The recommended formula looks as follows:

$$Model\ 2{:}\ SWF_c = SWF + 0.0058 * (T_{water} - 20) - 0.0154 \sin(\frac{2\pi}{365}(x + 4))$$

This model should correct the SWF value to measurements performed on the same place where the water temperature is 20 degrees Celsius at the end of June.

7.4. Which restriction should be set for a reliable model?

Currently, Rijskwaterstaat uses one restriction that does not allow measuring after a period of drought. This restriction is set on 14 days without more than 1 mm of rain on a day. One of their questions was to see if the restriction of drought could be extended.

In Chapter 6 the influence of drought has been determined. A problem for this analysis is that all the measurements are already performed with the current restriction. Therefore, it is not possible to calculate if there is a significant influence between the two groups. Thus, with the current data, we cannot answer if the *current restriction is sufficient,* directly. We did calculate if the restriction could be lowered.

We did conclude that there was a significant difference between the measurements after some periods of drought. However, the Pearson coefficient indicated a weak correlation. We would not recommend reducing the limit. Not only because of the low correlation it has with the SWF, but also the *trade-off between reliability and practicality*. This reduction would limit the number of days that measurements could be performed.

In the next part, we analysed the difference between the corrected SWF values of measurements with a period of drought (where the maximum amount to consider something dry is higher). This resulted in weak correlations with small p-values.

Therefore, we did **not** find any reason to believe that the *current restriction is **in**sufficient*.

To *increase the reliability of the model,* we would suggest additional researches. Research about the influence of tyre temperature, or research where water temperature and road surface temperature are not highly correlated, can help with formulating a new and maybe more reliable model. This could also be done by research that uses a lot more data or one that already knows the actuals SWF value and does not have to use Assumption 0. To increase the reliability of this model, we would recommend doing analysis with data where the measurement does **not** fulfil the current requirement.

8. References

1. E. Vos, T.B., F. Bouman, P. Kuijper, J. Voskuilen, J. Groenendijk (KOAC-NPC), Ministerie van Infrastructuur en Milieu, Rijkswaterstaat Grote Projecten en Onderhoud (RWS, GPO), *Skid resistance on national roads*, G. Utrecht : RWS, Editor. 2015.
2. Rijkswaterstaat. *Stroefheidsmetingen*. 12-08-2019; Available from: https://www.rijkswaterstaat.nl/zakelijk/werken-aan-infrastructuur/bouwrichtlijnen-infrastructuur/autosnelwegen/stroefheidsmetingen.aspx.
3. Marta Pagola, O.G., *Analysis of Temperature Influence on Skid Results.* international surface friction conference, Safer Road Surfaces, Saving Lives, 2011. **3**.
4. Baran, E., *Temperature Influence on Skid Resistance Measurement.* international surface friction conference, Safer Road Surfaces, Saving Lives, 2011. **3rd**.
5. Heerkens, H., & van Winden, A., *Solving Managerial Problems Systematically.* 2017: Noordhoff Uitgevers.
6. (PGBS), P. *Six steps in CRISP-DM the standard data mining process*. 2019; Available from: http://www.proglobalbusinesssolutions.com/six-steps-in-crisp-dm-the-standard-data-mining-process/.
7. KNMI, *Dagwaarden neerslagstations*, KNMI-neerslagstations, Editor.
8. Turban, E., et al., *Business intelligence : a managerial approach*. 2008, Upper Saddle River, N.J.: Pearson Prentice Hall.
9. Lu, Q., B. Steven, and F. No, *Friction testing of pavement preservation treatments: literature review.* Compare, 1971.
10. Henry, J.J. and B. Hill, *Short-term weather-related skid resistance variations.* Transportation Research Record, 1981. **836**: p. 76-82.
11. Rice, J., *Seasonal variations in pavement skid resistance.* Public Roads, 1977. **40**(4).
12. Ltd, L.R., *Pearson Product-Moment Correlation*. 2018.
13. McLeod, S.A., *What a p-value tells you about statistical significance.* Simply Psychology, 2019.
14. *How to Interpret Regression Analysis Results: P-values and Coefficients*, Minitab, Editor. 2013.
15. Glen, S., *Variance Inflation Factor*. 2015.
16. Team, G.L., *Linear Regression in Machine Learning*. 2020
17. Liu, C.-W., et al., *How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation.* arXiv preprint arXiv:1603.08023, 2016.
18. Kangas, M., M. Heikinheimo, and M. Hippi, *RoadSurf: a modelling system for predicting road weather and road surface conditions.* Meteorological Applications, 2015. **22**(3): p. 544-553.
19. Smith, K.W. and M.S. Sasaki, *Decreasing multicollinearity: A method for models with multiplicative functions.* Sociological Methods & Research, 1979. **8**(1): p. 35-56.
20. Mansfield, E.R. and B.P. Helms, *Detecting multicollinearity.* The American Statistician, 1982. **36**(3a): p. 158-160.

9. Appendix

9.1. Sinusoid

$$Sin\left(\frac{2\pi}{365}(x-\beta)\right) = Sin\left(\frac{2\pi}{365}x\right)\cdot Cos\left(\frac{2\pi}{365}\beta\right) - Cos\left(\frac{2\pi}{365}x\right)\cdot Sin(\frac{2\pi}{365}\beta)$$

In this formula the x represents the day number, which can differ for each measurement. The β represents the phase shift which is a constant as that does not change per day. α when determining the coefficients.

$$SWF_{diff} = \alpha_1 \cdot \sin\left(\frac{2\pi}{365}x\right) + \alpha_2 \cdot \cos\left(\frac{2\pi}{365}x\right)$$

Here x is the day number and SWF_{diff} the unexplained difference between the measured SWF and SWF at the standard day number, (the standard day number is dependent of the phase shift and one of the two dates where the outcome of the sinusoid is zero).

After determining two (different) values for the coefficient, the phase shift and actual coefficient can be calculated since $(\frac{2\pi}{365}\beta)$ is a constant value which is included in the coefficient as following:

$\alpha_1 = \alpha \cdot \cos\left(\frac{2\pi}{365}\beta\right)$ and $\alpha_2 = -\alpha \cdot sin(\frac{2\pi}{365}\beta)$ with this the actual coefficient (α) can be calculated using the $\alpha = \sqrt{\alpha_1^2 + \alpha_2^2}$ and the phase shift (β) with $\beta = \frac{365}{2\pi} \cdot \tan^{-1}(-1 \cdot \frac{\alpha_1}{\alpha_2})$.

9.2. Example results multiple linear regression

Categorical predictor coding (1; 0)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	8	1,09480	0,136850	476,90	0,000
Troad	1	0,00803	0,008026	27,97	0,000
Twater	1	0,01804	0,018039	62,86	0,000
sinbx	1	0,02204	0,022038	76,80	0,000
cosbx	1	0,00015	0,000153	0,53	0,467
Measurementplace	4	0,77115	0,192787	671,83	0,000
Error	295	0,08465	0,000287		
Total	303	1,17945			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,0169399	92,82%	92,63%	92,31%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0,67386	0,00830	81,23	0,000	
Troad	-0,001110	0,000210	-5,29	0,000	2,18

Twater	-0,004568	0,000576	-7,93	0,000	5,88
sinbx	0,01679	0,00192	8,76	0,000	2,42
cosbx	-0,00234	0,00321	-0,73	0,467	3,41
Measurementplace					
10	0,00806	0,00593	1,36	0,175	1,74
A	0,12549	0,00547	22,93	0,000	2,39
B	-0,04126	0,00451	-9,15	0,000	5,11
C	0,04293	0,00451	9,52	0,000	5,25

Regression Equation

Measurementplace			
1	SWF	=	0,67386 - 0,001110 Troad - 0,004568 Twater + 0,01679 sinbx - 0,00234 cosbx
10	SWF	=	0,68192 - 0,001110 Troad - 0,004568 Twater + 0,01679 sinbx - 0,00234 cosbx
A	SWF	=	0,79935 - 0,001110 Troad - 0,004568 Twater + 0,01679 sinbx - 0,00234 cosbx
B	SWF	=	0,63260 - 0,001110 Troad - 0,004568 Twater + 0,01679 sinbx - 0,00234 cosbx
C	SWF	=	0,71679 - 0,001110 Troad - 0,004568 Twater + 0,01679 sinbx - 0,00234 cosbx

Fits and Diagnostics for Unusual Observations

Obs	SWF	Fit	Resid	Std Resid	
2	0,73470	0,69900	0,03570	2,16	R
3	0,71884	0,68046	0,03838	2,37	R
13	0,70364	0,74292	-0,03928	-2,41	R
14	0,67117	0,70735	-0,03617	-2,19	R
146	0,70802	0,67398	0,03404	2,03	R
198	0,61497	0,65230	-0,03733	-2,23	R
210	0,60332	0,63960	-0,03629	-2,16	R
212	0,60611	0,63986	-0,03375	-2,00	R
245	0,54333	0,58354	-0,04021	-2,39	R
249	0,54835	0,58758	-0,03923	-2,33	R
288	0,54000	0,57675	-0,03675	-2,24	R

290	0,54000	0,57708	-0,03708	-2,29	R	
296	0,64400	0,60604	0,03796	2,33	R	
299	0,64000	0,63824	0,00176	0,11		X
301	0,54800	0,55232	-0,00432	-0,27		X
303	0,54500	0,53758	0,00742	0,46		X

R Large residual
X Unusual X