EFFECT OF THE SURROUNDING ON THE CONSTRUCTION COST OF DISTRICT HEATING NETWORKS AND SIMILAR INFRASTRUCTURES

Analysis based on Drinking Water and Natural Gas replacement projects

J. Mieras

TU Delft

Deltares

Effect of the Surrounding on the Construction Costs of District Heating Networks and Similar Infrastructures

Analysis based on Drinking Water and Natural Gas replacement projects

by

J. Mieras

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday February 4, 2021 at 11:00 AM.

Student number:	4374193	
Project duration:	March 3, 2020 – February 4, 2021	
Thesis committee:	Prof. dr. ir. I.W.M. Pothof,	TU Delft, supervisor
	Dr. ir. P.W. Heijnen,	TU Delft
	Prof.dr. J.P. van der Hoek,	TU Delft
	Ir. L. van der Most	Deltares

Preface

Dear reader,

Before you lies my master thesis. This master thesis is the final step in obtaining my master's degree in Sustainable Energy Technology (SET), and it concludes my academic pursuits at the TU Delft. For the past 11 months working on this thesis, I learned a lot, met some really nice people, and even though there were some ups and downs I overall had a good time. Although I have to say that I am happy to conclude the report and finally present the results to all of you. However, this thesis would never be like it is now without some amazing, much appreciated, help I received. Therefore I would like to take this opportunity to thank some people.

First of all, I would like to thank my colleagues from Deltares. They not only came up with the original research proposal, but also provided me with some useful insights, helped with the data gathering, provided (much appreciated) critical feedback, gave me some basic (GIS) training, and of course general support in the biweekly meetings. It is really a shame that unfortunately, most of my work had to happen from home, and most of our meetings were therefore online. However, I genuinely appreciated the couple of days that we were able to work together in the office and I think it is a shame that we did not get to know each other better than we did. So thank you Ivo Pothof, Lieke van der Most, Lieke Husken, and Amine Aboufirass your help is much appreciated.

Secondly, I would also like to thank Linda Kamp and Petra Heijnen for their willingness to cooperate with this research proposal. Which provided me with the opportunity to graduate in the field of (district) heating, which even though it is related to sustainable energy, is unfortunately not (yet) represented in the SET master. Furthermore, I would also like to thank Petra for her useful insights and challenging questions that I believe added value to the research.

In the end, for my research, I conducted interviews with 50 different people, every single one I would sincerely like to thank for their time, their willingness to share their knowledge, and their critical remarks. Out of these 50, the following deserve some extra attention. Jan Peter van der Hoek for his, much needed, speed course on drinking water infrastructures and his willingness to be a part of my graduation committee; Arne Bosch for sharing his contacts in the water sector and thereby saving me a lot of time; and some extra special thanks to Peter Horst, Mahsum Yilmaz, Lennard Wools, Sjoerd Buddingh, Corina van der Hulst, Ko Spruit, Berend Nootenboom, Bob Goessen, Arjan Hekker, Robert van den Brink, Peter van Houwelingen for their efforts in, and their willingness to cooperate with, the data gathering of this research.

Finally, I would like to thank my friends and family for their support. Especially, my brother, father, roommates, and study friends who took the time to check some parts of my report, and my lovely girlfriend who designed the front page and some other visuals in the report that look too good to be made by an energy engineer.

I hope you enjoy reading the report!

J. Mieras
Rotterdam, January 2021

Executive summary

The Dutch government set the goal to eliminate the use of natural gas in the residential sector by 2050. To attain this goal, new sustainable heating alternatives should be implemented on a large scale in the upcoming years. District Heating (DH) systems have the potential to play an important role in this energy transition towards a decarbonized energy system in the built environment.

However, accurate estimates of the cost of implementation of DH networks are important to realize its full potential. In this thesis the construction costs of DH are investigated. This is because current energy transition models (VESTA, ETM, CEGOIA, Comsof heat, etc) that calculate the costs for different sustainable heating solutions in the Netherlands, lack a detailed uniform overview of the construction costs of DH networks. The cost estimates of these models vary and or have large bandwidths. According to VESTA documentation, it is likely that an important reason for their large bandwidth is the fact that surrounding factors, like pavement type or ground pollution, are not included in their construction cost calculations. Reducing this bandwidth by including surrounding factors in construction cost predictions is the main goal of this research. Realizing this and thereby providing municipalities, policymakers, and heating companies with more accurate cost estimations is a valuable contribution to the Dutch energy transition in the built environment.

To include surrounding parameters in construction cost modeling historical project data is required. In the Netherlands, however, only approximately 6% of all the households was connected to a DH network in 2018 [35]. Combining this with the facts that some of the current DH networks are very old and therefore less relevant, and that it is unrealistic to assume that historical project data can be gathered from all potential DH projects. It can be concluded that gathering sufficient amounts of historical project data poses a serious challenge. That is why, in this thesis, historical project data from similar infrastructures (drinking water, natural gas) is used to analyze the effect of surrounding parameters on the construction costs of fluid infrastructures in the built environment. The construction of these types of infrastructures is relatively similar and therefore it is assumed these infrastructures have similar cost dependencies on surrounding parameters. Water and Gas are used in the analysis because these infrastructures, in contrast to DH networks, are present in (almost) the entire country.

It is important to note that originally heat data was supposed to be included in the analysis. However, unfortunately, no historical project data for heat projects was received in time to include in the research. Because of this, the results of this research, can not directly be used to predict the construction costs of DH networks. Nevertheless, the results of this analysis are off added value for DH prediction in the following three ways: First, the study can provide a good proof of concept for construction cost modeling considering surrounding parameters; Second, since drinking water, natural gas, and DH networks are relatively similar, it is quite likely that the significant relations found in the Water and Gas model are also applicable up to some extent for DH construction cost estimations. Third, the data gathered in this research can potentially be used as extra training data for a (future) DH heating model.

For this research data is gathered for drinking water and natural gas replacement projects in the built environment in the Netherlands. Three different types of inputs are required: costs, network properties, and surrounding parameters. Project costs and network properties are gathered using a data-gathering spreadsheet from water and gas utility companies. The final input type, surrounding parameters, are gathered from national open-source GIS databases. After talking to all water utilities and the three biggest network operators in the Netherlands data from 70 water projects and 52 gas projects was gathered and used as input for this research. Furthermore, data on 19 different surrounding parameters gathered from 12 different GIS databases is added to the costs and network properties of the 122 projects. These 19 potentially important surrounding parameters were identified in approximately 40 interviews with domain experts.

Before the modeling of construction costs was started a state of the art analysis on currently available cost calculation tools for DH networks was conducted. Thirteen different energy transition model were studied with a specific focus on how these models calculate the construction cost of DH networks. Three important conclusions can be drawn from this analysis. First, there is significant room for improvement regarding the level of detail of the cost calculations. Second, no or hardly any surrounding parameters are currently considered when predicting the construction costs. Finally, most (10/13) studied models offer the possibility to enter your own specific cost key figures.

In this research, a statistical model is developed, that generates detailed cost parameters (€/m), for (fluid) infrastructures in existing neighborhoods, considering a variety of surrounding factors that can influence the construction cost. It was decided to use Machine Learning (ML) to develop such a model. Therefore a literature review on different cost modeling techniques is conducted with the primary focus on ML approaches for costs prediction. Characteristics of different ML approaches were compared, and literature was studied in which ML was applied for similar purposes. Based on this literature review it was concluded that linear regression was the best suited ML approach. This conclusion was drawn based on three reasons: First, for similar cost prediction projects for DH and similar infrastructures linear regression is the most common ML approach to predict cost; Second, when data sets are small linear regression seems to work (at least relatively) better [56]; Third, the results of a Linear regression model are more transparent and therefore easier to interpret and compare with different (linear regression) models. This comparability is important since a comparison between construction cost of different infrastructures is required to attain the goal of this study.

Applying linear regression on the water and gas project databases to predict construction costs for water and gas infrastructures can be done in two different ways. The first approach is only using the water data to predict water construction costs and only using gas data to predict gas construction costs. The second approach is to combine the data sets to generate extra training data for both water and gas to potentially increase the model performance. It was concluded that using the water data as extra training data for the gas model did improve the model performance. However, using the gas data as extra training data for the water model did not increase the model performance. Based on this experience it is concluded that using the water and gas data as extra training data for heating predictions could increase the model performance but this is not conclusive and has to be validated in follow-up research.

In the research, an algorithm is developed which automatically improves the modeling performance of both the gas and water models by altering the modeling settings and input parameters. When this algorithm is applied to generate the best performing models it was found that, besides the average diameter, the pollution category (surrounding parameter) and 1/(total length) (no surrounding parameter) are the two most promising input parameters, for predicting construction costs of water and gas replacement projects. The pollution category strongly influences the costs since extra safety measures are required when the ground is polluted. Including 1/(total length) in the model compensates for the fact that small projects are relatively really expensive per meter because of the relatively big start-up costs. Including these, and other input parameters, in a linear regression model resulted in significantly better modeling performances. The best performing gas model has a $R2\_test$ score of 0.53, which is significantly better than the baseline model (only the diameter as input) $R2\_test$ score of 0.17. For the water model the $R2\_test$ score is increased from 0 (baseline) up to 0.43 (best performing water model). When these $R2\_test$ scores are converted to bandwidth reduction, it is concluded that a bandwidth reduction of 25% with respect to the chosen baseline (using only diameter) for both the water and gas model was achieved. Additionally, the Mean Squared Error (MSE) of the best performing water and gas models compared to the baseline models are respectively 34% and 35% smaller. The fact that both these scoring criteria show a serious increase in modeling performance strongly indicates that the used methodology has serious potential.

All in all, it is believed that even though there are some important differences between the three considered infrastructures, the methodology, and potentially the data, provided by this research can be used to increase the prediction accuracy of construction cost of DH networks. Increasing this prediction accuracy is a valuable contribution to the Dutch energy transition in the built environment, Because, the energy transition models that are used by municipalities, can implement these more accurate predictions. By increasing the accuracy of these energy transition models municipalities are able to maker better decisions when designing their transition vision heat. This will to a faster and potentially cheaper energy transition in the built environment in the Netherlands.

List of Figures

List of Tables

Nomenclature

ΔT Temperature difference between supply and return flow [C°]

λ Penalty weight term in L1, L2 and elastic net regularization

$\overline{L_{connection}}$ Average connection length households [m]

$\overline{L_{dis_road}}$ Average distance a house is removed from the street[m]

$A_{network}$ The area of the district heating network [m^2]

B_{type} The building type [-]

B_{year} The construction year of the buildings [-]

$bagging_{boolean}$ Whether or not bagging is applied in regression model

$boolean_{asphalt}$ Presence of tar in the asphalt layer boolean [-]

$CAPEX$ Total construction cost [€]

$CAPEX_{buffer}$ Construction cost heat buffer in network [€]

$CAPEX_{connection}$ Construction cost connection pipes [€]

$CAPEX_{distribution}$ Construction cost distribution network [€]

$CAPEX_{stations}$ Construction cost network stations [€]

$CAPEX_{transport}$ Construction cost transport network [€]

Ch_{buffer} Cost per household for heat buffer [$\frac{\text{€}}{\#}$]

$Ch_{rural/urban}$ Construction cost per household [$\frac{\text{€}}{\#}$]

Ch_{sweden} Construction cost per house [€/#]

$Civil_{cost}$ Cival engineering cost of a pipe segment [€]

$Cm_{distribution}$ Cost per meter distribution network [$\frac{\text{€}}{m}$]

$Cm_{pipe}(P)$ Cost per meter pipe as a function of peak load in corresponding pipe [$\frac{\text{€}}{m}$]

$Cm_{transport}$ Cost per meter transport network [€/#]

$Cost_{com}$ Cost for compensating inhabitants for discomfort [€]

$Cost_{man}$ Project management cost [€]

Cp_{HTS} Cost per kilowatt heat transfer stations [$\frac{\text{€}}{kW}$]

Cp_{sub} Cost per kilowatt sub stations [$\frac{\text{€}}{kW}$]

$Cp_{transport}$ Cost per kilowatt transport network [$\frac{\text{€}}{kW}$]

d_i diameter of pipe segment i [m]

f_{detour} Factor of extra distance a transport pipe needs with respect to a straight line connection [-]

f_{dis} Cost factor distribution pipes (default 1) [-]

f_{LT} Factor of extra distance a transport pipe needs with respect to a straight line connection [-]

f_{MT} Factor of extra distance a transport pipe needs with respect to a straight line connection [-]

$f_{stations}$ Cost factor heat transfer stations (default 1) [-]

f_{trans} Cost factor transport pipes (default 1) [-]

H_{trench} excavation depth [m]

$L1_{wt}$ Regularization term to set proportions in elastic net regularization

L_i Length of pipe segment i [m]

$L_{dis\_network}$ Length of the distribution network [m]

l_{house} Pipe length per house [m/#]

L_{road} Total length of the road in the network area [m]

L_{source} Distance from (primary) heat source to neighborhood [m]

$L_{trans\_network}$ Length of the transport network [m]

$m_{bagging}$ Amount of models used in bagging algorithm

$Mechanical_{cost}$ Mechanical engineering cost of a pipe segment [€]

$N_{connections}$ Number of households connected to single connection pipe [#]

$N_{dis_{s}eg}$ Number of distribution pipe segments [#]

$N_{households}$ Number of households connected [#]

$N_{pipe_{s}egment}$ Number of pipe segments considered in calculation [#]

$N_{stations}$ Number of heat transfer stations [#]

$N_{trans_{s}eg}$ Number of transport pipes with different diameters [#]

$P_{connection}$ Peak load in single connection pipe [kW]

P_{peak} Peak load in corresponding pipe [kW]

S_{type} Surrounding type [-]

Tb Temperature building [$C°$]

Acronyms

R^2 Coefficient of determination. 28, 97, 101–103, 108

4GDH 4th Generation District Heating. 6, 7, 32

5GDH 5th Generation District Heating. 6, 7

CAPEX Capital Expenditures. 7, 8, 10, 11, 15, 16, 19

CHP Combined Heat and Power. 4, 5, 7

DH District Heating. 4–12, 15, 16, 19–22, 24, 32–35, 37–40, 43–46, 48, 50, 52, 53, 61, 77, 91, 93–95, 97, 98, 105–110, 115

DHW Domestic Hot Water. 5

EBN Energie Beheer Nederland. 3

ETA Energie Transitie Atlas. 16

ETM Energie Transitie Model. 16

GAMs Generalized Additive Models. 30

GBTs Gradient Boosted Trees. 30

GIGO Garbage in Garbage out. 16

HT High Temperature. 4, 15

HTS Heat Transfer Station. 4, 16, 19, 52

LT Low Temperature. 4, 7, 12

MAPE Mean Absolute Percentage Error. 28, 29

ML Machine Learning. 15, 24–31, 34, 35, 37, 38, 40, 46, 54, 56, 61–63, 65, 67, 68, 70, 71, 86, 88, 99–101, 105, 106, 108, 109, 115, 116

MSE Mean Squared Error. 28, 29, 65, 67

MSE Mean Squared Error. 97, 101–103, 108

MT Medium Temperature. 4, 5, 12

NN Neural Network. 30, 31, 34

NRMSE Normalized Root Mean Square Error. 28, 29

OPEX Operating Expenditures. 7, 8, 10

PBL Netherlands Environmental Assessment Agency. 9

RVO Rijksdienst Voor Ondernemend Nederland. 3

SVR Support Vector Regression. 30

VeWa Veiligheidsvoorschrift Warmte. 4

VLT Very Low Temperature. 4, 7

Contents

I

Part 1 Research outline

1 Introduction

In a reaction to counteract global climate change, close to 190 countries came together in December 2015 to sign the Paris Agreement. They agreed to cut down global emissions to reduce global warming. As a result, the member of the EU pledged to reduce emissions by 40% relative to the year 1990 by the year 2030[15]. The Netherlands currently one of the biggest emitters of the European Union has a long way to go. In 2017 according to CBS, the Netherlands accounted for 4.5% of Europe's total emissions resulting in a per capita pollution which is 34% higher than the European average[9]. It should be noted that when looking at the energy usage divided by the GDP for all European countries that the Netherlands is performing significantly better. Meaning that the high emissions can partially be explained by the relatively high production rates in the Netherlands. The Netherlands does however want to reduce the European total emissions to at-least 80% by 2050 [17] and the emissions reduction goals in the Netherlands itself are even higher. In "het regeerakkoord" of 2017, it is stated that the emissions in the Netherlands should be reduced to 49% in 2030 and 95% in 2050 relative to the 1990 emissions [18]. These extreme reductions should be realized by a serious decrease in fossil fuel usage in the Netherlands.

According to Energie Beheer Nederland (EBN) 90% of the total dutch energy supply is currently provided by fossil fuels whereas only 7% is renewable, which is the lowest of all European member states [2]. Of the fossil fuel supply, the biggest share (41%) is natural gas. This natural gas is for approximately 90% produced in the Netherlands itself [30]. A serious reduction of this natural gas production would thus be necessary to realize the set climate goals. This together with the serious earthquake issues caused by the natural gas production in Groningen led the Dutch government to make plans to drastically reduce the gas production in the coming years. Eventually to the point where, in 2030 the gas production must come to a full stop[11]. Removing such a big share from our national energy supply does not come cheap and causes some challenges. Maybe the biggest of these challenges are faced in the heating sector currently taking account of 41% of the total energy demand. Especially in the residential sector, a solution is needed. Currently, 10 billion m^3 of natural gas (28% of all the natural gas) is used to heat houses every year, accounting for 64% of total yearly energy usage of a house[30].

To achieve the climate goals and reduce the natural gas demand the dutch government set another goal. They decided to heat all houses and building without natural gas in 2050. This means that starting from 2021 every year 200.000 households should be transformed into an alternative, sustainable, heating solution [18]. According to a research conducted by Nieman Raadgevende Engineers and commissioned by Rijksdienst Voor Ondernemend Nederland (RVO) the three main solutions to heat houses in the Netherlands without natural gas are[59]:

- All electric
- District Heating networks
- Sustainable gas (green gas)

When comparing these solutions it is important to note that even though green gas is currently the cheapest natural gas alternative it will probably not become the main solution. This is due to the lack of available green gas. It is estimated that only 10% of all the buildings in the Netherlands can be heated using green gas in 2050 [59]. When comparing All-electric to District Heating (DH) networks in a very generic way, a decision must be made between high cost for the collective (DH) and high costs for individual homeowners (all electric). In general, it is assumed that DH networks are a better solution for areas with a high energy demand density $kWh/year/m^2$ and all-electric is more suitable for areas with a low energy demand density. However, the prices, and thus the preferred solution, are also very much dependent on the characteristics of the buildings and the surrounding in a certain neighborhood (road types, building type, insulation level etc.) and therefore a more in-depth analysis is recommended when choosing a preferred solution.

1.1. District Heating networks

As mentioned above it is likely that different neighborhoods require different solutions and that a more in-depth analysis is preferable for choosing the “best” solution. For this reason, doing a more in-depth analysis in any of the three above mentioned solutions is of added value for the Dutch energy transition in the built environment. Out of the three listed sustainable heating solutions, this research only focuses on the second solution, DH networks. This is not necessarily because it is believed that DH is a superior solution compared to the other two. This decision is made because, as also explained in section 1.6, the scope of track two of the WarmingUP consortium, which this research is a part of, is researching DH potential. In this subsection, some background information on DH networks is presented. First, in subsection 1.1.1 the working principle of heating networks is briefly explained. Second, subsection 1.1.2 describes the development of DH networks over time giving a short overview of the different generations of DH networks. Finally, the potential and challenges faced by DH networks are explained in subsection 1.1.3.

1.1.1. Working principle and definitions

DH networks are closed double looped pipe systems that connect buildings to heat sources. They consist of two pipelines a supply and return pipe. The supply pipe transports hot water to buildings and the return pipes transports cold water back to the sources to be heated again. Since the system is closed-loop all the buildings are connected through a heat exchanger which extracts the heat from the network so it can be used for heating or taking showers. According to a definition document drafted by the heating companies participating in the WarmingUP consortium, DH networks can be categorized into four types. These four types can be distinguished based on the temperature delivered to the buildings (Tb) and are called: Very Low Temperature (VLT), Low Temperature (LT), Medium Temperature (MT) and High Temperature (HT) the temperatures corresponding to these types can be seen in Table 1.1.

Table 1.1: Different type’s of District Heating networks [61]

Type	Temperature
Very Low Temperature (VLT)	Tb <25[deg]
Low Temperature (LT)	25 <Tb <55[deg]
Medium Temperature (MT)	55 <Tb <70[deg]
High Temperature (HT)	Tb >70[deg]

In Figure 1.1 a graphical overview of a standard district heating network in the Netherlands, as defined by the energy companies taking part in the WarmingUP consortium, is shown [61]. Similar to drinking water and natural gas networks DH networks consist of a transport and distribution net depicted by numbers 1 and 3 in Figure 1.1 respectively. The transport net transports heat from the source, nowadays most of the time a Combined Heat and Power (CHP) plant [64], to a neighborhood supplied by a DH network. However, in the future, other more sustainable heat sources, like geothermal heat, and centralized heat pump solutions that use electricity, will become more common. In the neighborhood itself, the distribution network will transport the warm water to the street level. The distribution and transport network are connected in a so-called Heat Transfer Station (HTS), depicted by number 2 in Figure 1.1, a HTS is normally equipped with a peak demand boiler to boost the temperature of the main source during peak demand. Depending on the size of a district heating network and the temperature of the main source the distribution network can be divided into the primary distribution net (primary net) and the secondary distribution net (secondary net). The transport net and the primary distribution net ordinarily have higher temperatures and are therefore considered HT according to Veiligheidsvoorschrift Warmte (VeWa). The secondary distribution net typically has a lower temperature and is considered MT. In the VeWa safety standards described in [41] a part of the net is considered HT when the water temperature in the pipes (note the difference with Table 1.1 where Tb is delivered temperature) is above 100[deg] and MT when it is below 100[deg]. The two different distribution nets are connected with a substation depicted by number 4 in Figure 1.1. The last two numbers 5 and 6 in Figure 1.1 respectively represent a heat delivery station and heat delivery set. The difference is that a heat delivery set is smaller and only delivers heat to single households or small businesses. Where the heat delivery station can provide warm water to multiple households or companies in for example an apartment building or shopping mall.

The last type of pipe which is not depicted in the graphical overview is the connection pipe. The connection pipe forms the connection between the distribution network and the heat delivery sets and stations. The costs of this connection pipe are covered by the connection cost fee costumers have to pay once to connect their buildings to the heat network. Currently, the maximum connection fee that customers have to pay is determined every year by the Dutch government. This maximum price is fixed unless the distance to the distribution network for houses is more than 25 meters [7].

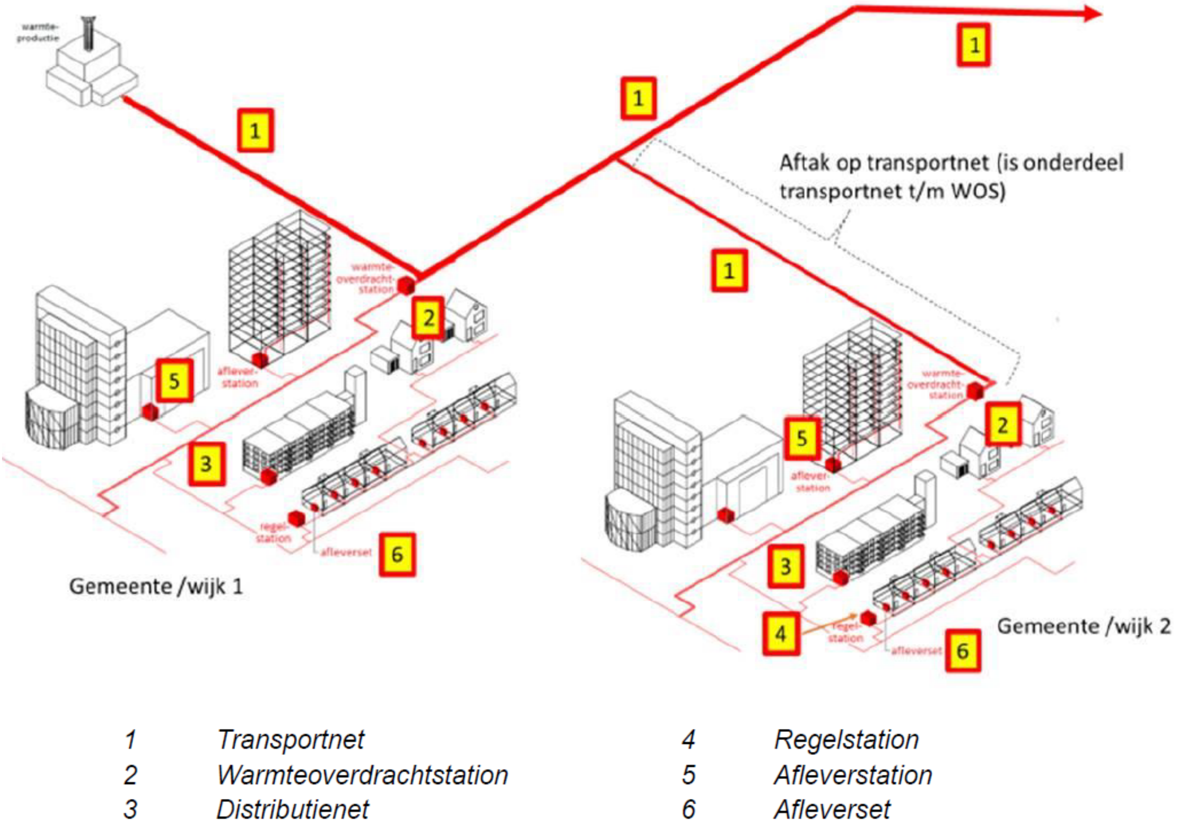


Figure 1.1: District heating network graphical overview constructed by consortium partners [61]

1.1.2. Different generations of district heating

Overtime DH networks changed significantly transforming to safer, more efficient, and more sustainable systems. According to Lund et l. four different generations of DH networks can be distinguished. In their 2014 paper, a definition of fourth-generation DH is defined [43]. The definitions provided in this paper are widely accepted and used in literature. A short description of the four different DH generations as discussed in the paper is given below.

In the 1880s the first DH networks were introduced in the united states. These first-generation networks used steam as a heat carrier. This steam was used directly in high-temperature radiators to heat houses and could be used to heat water tanks for Domestic Hot Water (DHW) use. The big disadvantage of such systems was the high energy losses and the risk of steam explosions killing pedestrians. This type of DH with steam as a primary heat carrier nowadays is only still used in Manhattan and Paris.

In the 1930s the second generation, of DH networks, using pressurized hot water, surfaced. This type of DH was mainly used to utilize heat from CHP plants. Even though the technology is outdated the original pipes from these networks are still used for the higher temperature parts of some older networks.

The third generation of DH networks emerged in the 1970s, these systems still used (pressurized) hot water as the main heat carrier but at lower temperatures now typically below 100[deg]. In third-generation systems solar and geothermal heat was used occasionally as heat sources for the first time also floor heating systems instead of conventional (high) temperature radiators were used on some occasions.

A clear trend towards lower supply temperatures and more energy-efficient (arguably more sustainable) systems can be identified in the DH sector. Logically the definition of 4th Generation District Heating (4GDH) follows this trend. In the paper from Lund et al., it is identified that for DH networks to successfully fulfill a role in a future (sustainable) energy system 5 challenges have to be faced. These five challenges are depicted in Figure 1.2. The definition of 4GDH provided in the article is given in the text box below.

4GDH: *The 4th Generation District Heating (4GDH) system is consequently defined as a coherent technological and institutional concept, which by means of smart thermal grids assists the appropriate development of sustainable energy systems. 4GDH systems provide the heat supply of low-energy buildings with low grid losses in a way in which the use of low-temperature heat sources is integrated with the operation of smart energy systems. The concept in- volves the development of an institutional and organisa- tional framework to facilitate suitable cost and motivation structures. [43]*

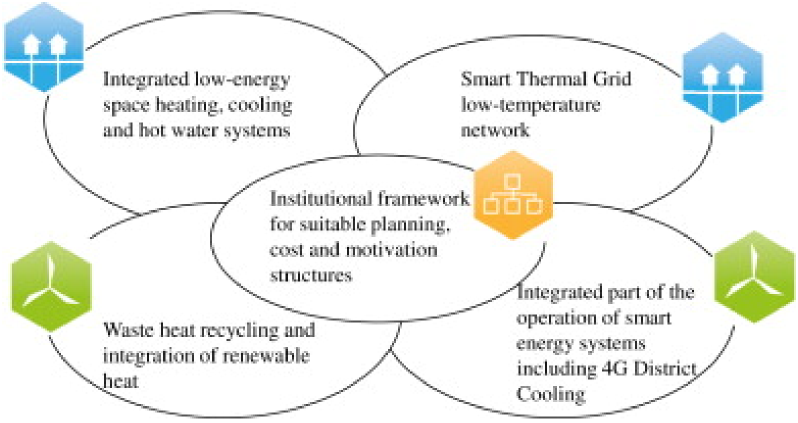

Integrated low-energy space heating, cooling and hot water systems
Smart Thermal Grid low-temperature network
Institutional framework for suitable planning, cost and motivation structures
Waste heat recycling and integration of renewable heat
Integrated part of the operation of smart energy systems including 4G District Cooling

Figure 1.2: The concept of 4th generation district heating [43]

According to Buffa et al. the 4GDH described by Lund et al. is not the ideal DH system to full-fill a successful role in the future energy system. In their paper, they propose a 5th Generation District Heating (5GDH) network which could outperform 4GDH. However, the goal of their paper is not to compete with the 4GDH concept but to "revise the definitions they encountered" [23]. When studying literature they discovered that multiple papers suggested "upgrades" or additions on the 4GDH concept. They found that many of the papers suggested a similar concept which they argue is not entirely in line with the concept of 4GDH. The two main differences between the newly proposed concepts (in the paper defined as 5GDH) and 4GDH are the following:

- The same pipes should be used for Heating and Cooling demand

- Supply temperatures should be even lower (close to ground temperature) to further decrease losses, this low temperature is compensated by booster heat pumps in the substations.

Based on the described concepts in literature and realized projects in for example Germany and Switzerland a definition of 5GDH is drafted. The definition given in the paper is given in the text box below.

5GDH: *A 5GDHC network is a thermal energy supply grid that uses water or brine as a carrier medium and hybrid substations with Water Source Heat Pumps (WSHP). It operates at temperatures so close to the ground that it is not suitable for direct heating purpose. The low tem- perature of the carrier medium gives the opportunity to exploit directly industrial and urban excess heat and the use of renewable heat sources at low thermal exergy content. The possibility to reverse the operation of the customer substations permits to cover simultaneously and with the same pipelines both the heating and cooling demands of different buildings. Through hybrid substations, 5GDHC technology enhances sector coupling of thermal, electrical and gas grids in a decentralised smart energy system. [23]*

It is good to indicate that a parallel can be drawn between 4th and 5th generation networks and the LT and VLT networks defined by the partner heating companies of the WarmingUP consortium, as discussed in subsection 1.1.1. Even though these are not directly linked it is safe to assume that the heating companies based their definitions of network types on the existing literature discussed above.

1.1.3. Potential and challenges of District Heating

Heat Roadmap Europe [27] and other recent studies see lot of potential in DH for sustainable future energy systems [40][20][39][19]. However, in 2015, in Europe, only 12% of the residential and services sector is connected to DH [31]. In the Netherlands in 2019 approximately 400.000 households were connected to DH [22] accounting for a little bit less than 6% of the total amount of houses. However with the current climate policy and the heating without natural gas goal in the Netherlands, the amount of houses connected to district heating is likely to increase significantly in the coming decades. In the masterplan aardwarmte, it is estimated that 40% of the heat demand in the Netherlands will be delivered through DH by 2050 [51].

Even though this transition from natural gas towards DH has a lot of potential to decarboinze the heating sector it is not necessarily true that district heating is sustainable. In 2017 only 27% of all the heat supplied to European district heating networks was sustainable [64]. Meaning that the rest of the produced heat still causes harm-full pollution and negative climate effects. It is important to note that 72% of the remaining 73% is so-called recycled heat meaning that the heat is a waste product from for example industries or electricity production (CHP)[64]. This heat would be lost if not used for DH purposes. Due to this phenomenon, the fundamental idea of DH incorporates energy efficiency and resource synergy [50]. But when considering that in the future the industries and the electricity production should also transform to more sustainable alternatives, it is clear that more alternative sustainable heating sources are required to provide heat to future DH networks. Meaning that the 4GDH and 5GDH systems discussed above probably have a serious role to play in the energy transition in the Netherlands and all over Europe.

The design and investment process of a DH network is complicated. This is due to the long lifetime of the network (more than 40 years) and the trade-off between high upfront Capital Expenditures (CAPEX) and lower yearly operational cost (Operating Expenditures (OPEX)) [44]. "Better" networks require higher upfront investments but need less fuel because of lower heat losses and less pumping requirements. Also, better-designed networks require less maintenance and therefore have lower maintenance costs. Two factors that significantly influence the design and investment process of DH networks are surrounding parameters and the network temperature.

Both these factors therefore also influence the investment cost. However, the different design temperatures of a network are researched extensively, resulting in the different generations of DH networks discussed in the previous section. As pointed out the trend is to reduce design temperature levels because networks need to become more sustainable in the coming decades. These lower temperatures are required because they will reduce costs when using sustainable heat sources. In a recent study by Averfalk and Werner, it was found that lower supply temperatures have a significant positive influence on cost when using geothermal heat, industrial heat, industrial excess heat, and heat pumps. However, when using CHP plants running on either natural gas or biomass there is almost no difference in cost when the temperature is reduced [19]. This is mainly due to higher efficiency of the mentioned sustainable heat sources at lower temperatures. Also, some sustainable sources are not able to produce high temperatures and therefore need more peak booster capacity to adhere to the required temperature levels.

However, the relation between surrounding parameters and the CAPEX of district heating networks, is a relatively unknown research field. Resulting in uncertainty in the investment cost of new networks, when these networks are constructed in new unknown areas. This uncertainty causes problems when designing new district heating networks (trade-off between CAPEX and OPEX). But also when choosing between the three potential heating solutions mentioned in chapter 1 uncertainty on investment cost is undesirable. Since this increases the risk, of choosing the "wrong" solution, or choosing the "right" solutions but implementing it in the "wrong" way.

1.2. Problem formulation

After giving a short introduction of the context it is now time to specify the problem that is being looked into in this research. As mentioned, the Dutch government wants to reduce natural gas usage in the residential sector. To realize this goal, new sustainable heating alternatives should be implemented on a large scale in the upcoming years. A potential solution for this is implementing DH networks. However to implement DH in an efficient way it is important to be able to accurately predict investment costs in an early design stage. This is important because not all neighborhoods in the Netherlands are equally suited for implementing DH networks. When the less suited neighborhoods are chosen first, the projects are likely to suffer big losses and the changes of DH networks reaching their full potential decrease since municipalities are more likely to choose alternatives when facing similar choices in the future.

Currently, municipalities in The Netherlands have the responsibility to choose a sustainable heating solution for all their neighborhoods. In their transition vision heat, they have to provide a timeline indicating which neighborhoods should make the transition first, which neighborhoods should be last and which sustainable heating solutions should be used [6]. For municipalities to make these decisions without proper insight into the accompanying costs of the different alternatives is undesirable. Municipalities, especially smaller ones, are unlikely to have the required knowledge to choose the most suited neighborhoods with the "best" heating solutions. That is why it is very important to provide these municipalities with the right tools to assist their decision-making process.

An important criteria in the decision-making process are the costs of the different solutions. The costs for realizing new sustainable (heating) infrastructures can be subdivided into two groups the CAPEX and OPEX. Where CAPEX represents the upfront investment costs and OPEX the operation and maintenance cost. Both these costs are important when comparing different heating solutions. However, in this thesis, only the investment costs of DH are researched. To be more precise only the construction cost of the network itself is considered.

Of course, the construction cost of implementing the different heating alternatives shouldn't be the only criteria in choosing a suitable heating solution for a certain neighborhood. The diagram in Figure 1.3 shows how the construction costs fit in the bigger pictures of choosing the best solution. It can be seen that multiple other aspects should also be considered and tools to support municipalities with these decisions are also important but are outside of the scope of this research.

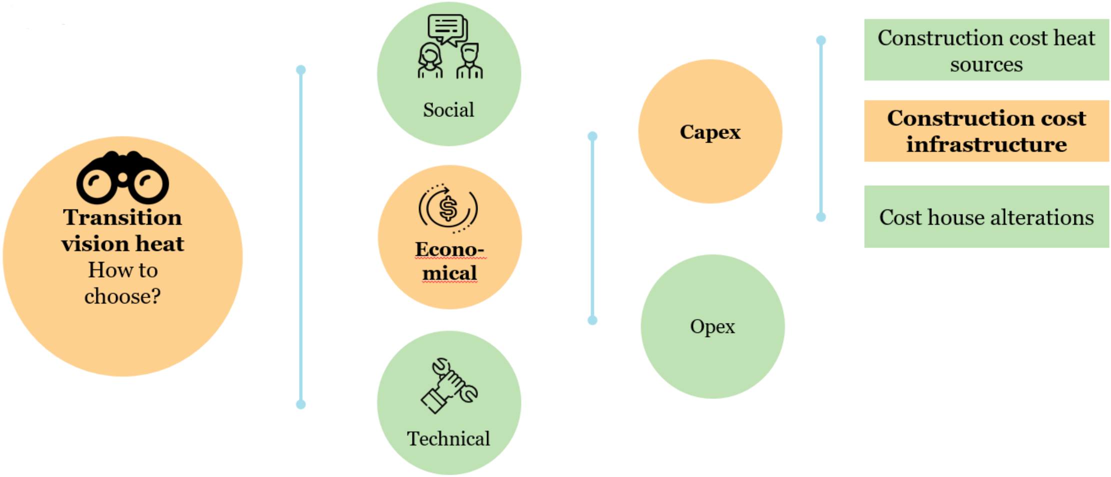


Figure 1.3: Overview of aspects that a municipality should consider choosing sustainable heating solution for a certain neighbourhood

To support municipalities and other policymakers on the cost aspect of choosing the "right" heating solutions multiple so-called energy transition models are developed. These models calculate the costs of different heating solutions. Most of these models calculate both the CAPEX and OPEX of the different heating sustainable possibilities. Although, it is important to note that these cost calculations are rough predictions since at this stage no detailed designs for the different heating solutions are available. Making these detailed designs is expensive and time-consuming, which is undesirable when a municipality is still in the decision-making process and might still choose a different solution. Therefore, it is preferable to first choose a solution based on a rough cost estimation before making detailed designs and contacting contractors.

Even though, the energy transition models are meant to make rough cost estimations, it is desirable that these estimations are as accurate as possible. If estimations are more accurate, the above-described risk of picking the wrong solution for a certain neighborhood decreases and the energy transition will happen more smoothly. However, the bandwidth of the available energy transition models is currently relatively large, which results in unaccurate predictions. This conclusion was drawn by the WarmingUp consortium after talking to stakeholders and is the main reason behind the research in theme 2C of the consortium. This thesis is meant as a contribution to the research in theme 2C. More information on the WarmingUP consortium can be found in section 1.6.

When for example looking into VESTA, the energy transition model developed by Netherlands Environmental Assessment Agency (PBL), which is also the model that is recommended by the government as a supporting tool for the transition vision heat. It can be found in their documentation that it is likely that the main cause of their large bandwidth is the fact that surrounding factors like pavement type or ground pollution, are not included in their calculation. The fact that VESTA developers knew that excluding the surrounding parameters from their model is likely to increase their bandwidth. But they still decided not to include them, strongly indicates that no sufficient information on surrounding parameters is currently available in the Netherlands. Since these surrounding parameters can vary significantly construction cost of heating networks with very similar dimensions can still be significantly different. For example, the cost of a new DH network in a newly built neighborhood are a lot lower than the cost of a new network in the city center of Amsterdam.

In general cost predictions in open field conditions (newly built neighborhoods) where everything is very predictable is relatively easy. However, cost predictions in densely populated areas with a lot of unknown factors are a lot harder. Meaning that models using cost parameters, which are solely dependent on the dimensions of the network (€/meter, €/kW, €/household, etc) without considering surrounding parameters, are having a hard time making accurate cost predictions in these densely populated areas. This is a problem since highly populated areas turn out to be the most promising areas for DH networks because of their high head demand densities. This implies that including surrounding factors in the cost prediction models and thereby providing municipalities, policy writers, and heating companies with more accurate cost estimations is a valuable contribution to the Dutch energy transition in the built environment.

To include surrounding parameters in construction cost modeling historical project data is required. In the Netherlands in 2018 only approximately 6% of all the households is connected to a DH network [35]. The other 94% is still connected to the natural gas grid. Also, some of the DH networks that are currently in existence in the Netherlands are relatively old. For example, the biggest DH network in the Netherlands, the Eneco network in Utrecht, exists since 1923 [12]. The construction costs of these older networks are not representing the current market situation properly. This means that the total amount of DH projects, that in theory could be available for this analysis, is limited. Moreover, probably not all the companies that realized these projects are willing to share data and/or saved the right data in their computer systems. This implies that, when doing an analysis using only historical project data from DH networks, gathering enough data to get statistically sound results could potentially be a serious challenge. Furthermore, since only 6% of the houses is currently connected to DH networks. It is likely that certain surrounding conditions, that could potentially influence the construction cost, are not present in any of these historical projects. Meaning that the developed model is only valid for certain parts of the country. For these two reasons, the WarmingUP consortium decided it might prove useful to include similar network infrastructures to expand the data base for DH cost predictions. This thesis resulted from this decision since it was decided that analyzing the potential added value of similar infrastructures for including surrounding parameters in DH cost predictions would be a perfect thesis topic.

1.3. Research objective

The main goal of this research is to develop a statistical model that generates detailed cost parameters (€/m) for DH networks in existing neighborhoods considering a variety of surrounding factors that can influence the construction cost. However as already mentioned in section 1.2 using only DH project data as input for a statistical model is a risk because of the potential lack of available data. That is why originally the objective of this thesis was to use historical project data from not only DH projects but also use historical data from similar infrastructures like for example drinking water, natural gas and sewer systems. The construction process of these infrastructures is relatively similar and therefore it is assumed that these infrastructures have similar cost dependencies on surrounding parameters. Furthermore, these infrastructures, in contrast to DH networks, are present in (almost) the entire country. This means that the total amount of potential projects is a lot higher and they have a bigger chance of generating a cost parameter model that is valid for the entire country.

Before the above described statistical model is developed some other related aspects are studied. Even though these other study topics are not the main goal of this thesis, they are also part of the total research objective. These studies are conducted to support the main research goal but can also be of added value as separate conclusions for other studies. First, the state of the art regarding DH cost calculations, specifically focusing on the construction cost calculations, is analyzed. The results of this analysis should show whether current models use surrounding parameters for their cost predictions and if indeed there is potential room for improvement by including them. The second aspect which is studied is the similarity of different infrastructures in the built environment. The results of this analysis should lead to a list of infrastructures that are deemed similar enough to use as (extra) inputs for cost predictions of DH networks. The third and final aspect that is studied, before the statistical model is developed, are the surrounding parameters that can potentially influence the construction costs. Besides looking into which surrounding parameters have the potential to influence the construction costs also potential (national) databases containing these parameters should be identified. This should result in a list containing potential surrounding parameter inputs for the statistical model.

1.4. Research questions

The research question of this thesis is:

Does including surrounding parameters in a statistical model, calculating construction costs of pipe infrastructures in the built environment, improve the model performance, and can a trained model and or project data from similar infrastructures be used to increase the model performance of a district heating model?

This research question let to the following subquestions:

1. *How do current energy transition models calculate construction cost for district heating networks?*
2. *Which modeling approach should be used to develop a model that calculates the construction costs of pipe infrastructures in the built environment?*
3. *Which infrastructures are similar enough to district heating networks to include in a combined cost modeling approach and what are the main similarities and differences between these infrastructures and DH networks?*
4. *Which surrounding parameters are potentially related to the construction costs of DH networks and similar infrastructures and where can (national) GIS databases containing this data be found?*
5. *Which surrounding parameters have the biggest influence on the construction cost of pipe networks in the built environment?*
6. *How can historical project data and resulting models for similar infrastructures best be used to improve a construction cost model for district heating networks?*

1.5. Scope and design of research

The goal of the research is to increase the accuracy of construction cost predictions in the built environment (€/m) by including surrounding parameters in the model. However, to realize this, it is important to define the terms, construction cost, and built environment. In this section, these two terms are explained and the scope of the research is defined. First, the costs that are included in the construction cost are discussed. The first logical distinction in cost is that only CAPEX are considered for the construction costs. Meaning that all the OPEX, like for example maintenance costs or pumping cost (energy use for pumping), are excluded.

When looking at the CAPEX of DH networks all the CAPEX that are related to the installation of pipes and substations are considered. That for example means that the CAPEX for construction of heat sources are excluded, but costs for permits and licenses (related to digging and installation of pipes) are included. However, not all the total cost (CAPEX related to pipes and stations), are necessarily included in the final construction costs per meter. Some costs, like for example material and design costs, are less likely to be correlated with surrounding parameters and for that reason, it is considered to subtract these costs from the total costs to improve the modeling performance. Whether these costs can and should be subtracted from the total costs is dependent on the availability of these specific costs in costs databases of data suppliers and the modeling performances after these costs are subtracted. More information on the subtraction of these costs can be found in section 5.5.

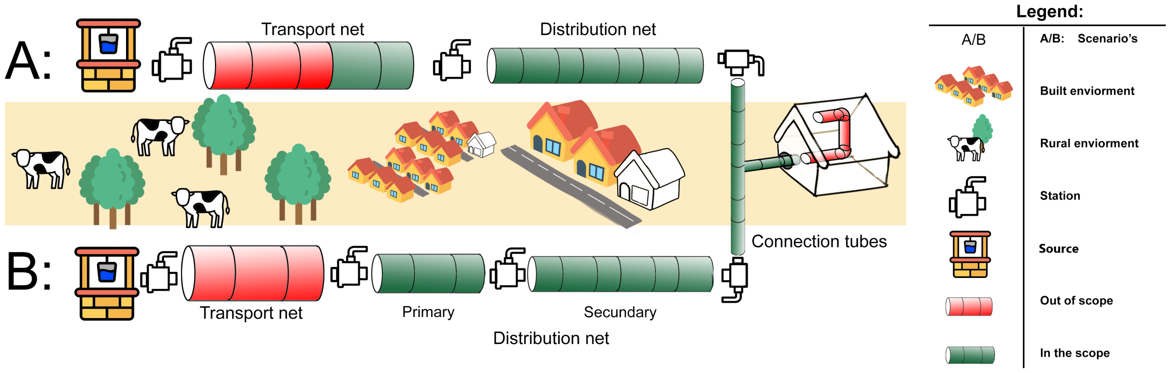


Figure 1.4: Research scope overview of the pipe segments included in the research

The second scoping requirement is the types of projects that are considered in this analysis. As mentioned above only projects in the built environment are considered. First of it is important to state that this distinction is not made based on network type. Meaning that for example all the transport pipes are excluded and all the distribution and connection pipes are included. This decision is made because different data providing companies might have different definitions for what they call transport and distribution pipes. Especially when an analysis is conducted comparing multiple different infrastructure types this is quite likely.

For bigger networks sometimes parts of the considered network (most of the time transport pipes) are outside of the built environment and this part is therefore not considered in the analyses. As mentioned the connection pipes of the network are considered in the analysis, but only up to 3 meters behind the facade of a building. Meaning that for example, indoor pipe systems in big apartment buildings are outside of the scope. Besides the pipes, all the (sub)stations for the different networks that are located in the built environment are also considered in the analyses. For a graphical overview of two different scenarios (A and B) see Figure 1.4. In the image, all the red pipes are excluded from the analyses, as can be seen sometimes that could mean that just a part of the transport pipes is considered. Also, the connection set and everything related to the production of heat, gas, and water are excluded from the analysis (geothermal power plants, (gas)boilers, water treatment plants, etc).

As can be seen in the image the border between the environment and the rural environment is no clear cut meaning that there is some room for interpretation. This represents the real-life situation where it might sometimes be hard to pinpoint the exact location where the built environment begins. The interpretation of which parts of the network are in and which parts of the network are outside of the built environment is left for the companies providing the historical cost data. However, data providing companies are advised to only provide projects which are entirely constructed in the built environment to prevent errors in the splitting of project data. It is chosen to initially give the responsibility for this scoping process to the data providing companies because it is assumed that they are the best suited to indicate which parts of the network faced limiting surrounding factors from the built environment. However, projects that were provided by the companies but did not have any houses within 30 meters of the constructed pipes were removed by hand. More information on the gathering of project data can be found in section 4.2.

1.6. WarmingUP

This thesis is part of the WarmingUp research consortium. In this subsection, a brief description of the WarmingUp consortium is given. This is done because the original idea of the thesis is part of the project plan of the consortium and the results of the thesis, if successful, are supposed to be used as a starting point for further research. Also, the consortium is involved in the data gathering process, and the historical cost data from heat projects was supposed to be provided by the project partners.

The WarmingUp innovation plan consists of 32 research projects subdivided into 6 cohesive themes. The goal of the research consortium is to gather social and technical based knowledge to design practical tools that can be used to realize cost-effective, socially acceptable, sustainable, and trust-worthy collective heating solutions in the built environment.

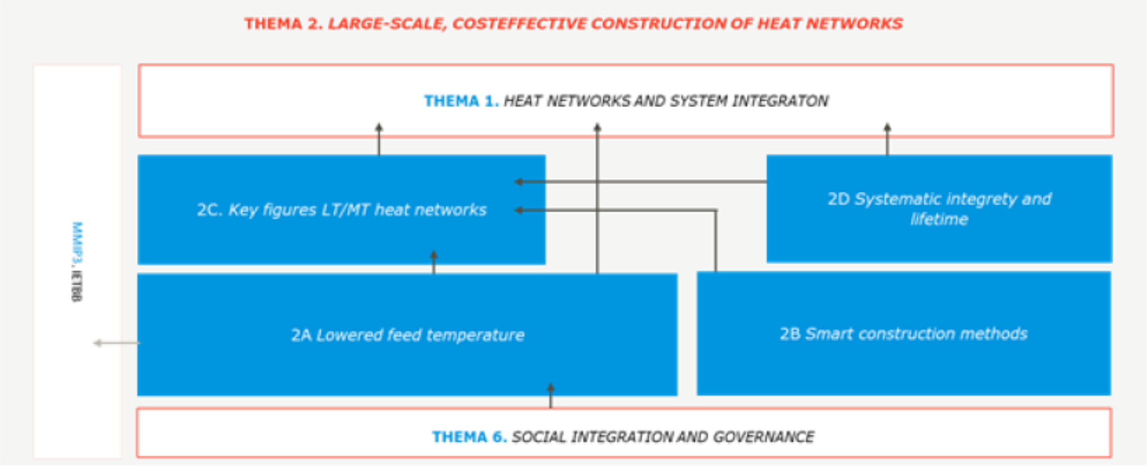

THEMA 2. LARGE-SCALE, COSTEFFECTIVE CONSTRUCTION OF HEAT NETWORKS
THEMA 1. HEAT NETWORKS AND SYSTEM INTEGRATON
2C. Key figures LT/MT heat networks
2D Systematic integrety and lifetime
MMIP3. IETB8
2A Lowered feed temperature
2B Smart construction methods
THEMA 6. SOCIAL INTEGRATION AND GOVERNANCE

Figure 1.5: Overview and connections WarmingUP theme 2 [62]

Theme 2, which this thesis is a part of, focuses on: "*Large-scale, cost-effective construction of heat networks*. The goal of this theme is to develop methods to support the scaling up of DH networks in the environment. Which should lead to 80.000 extra houses that are being connected to DH networks every year. Theme 2 is divided into four different research projects which are presented in Figure 1.5. As can be seen in the figure the different research projects are connected and are dependent on results from one another. The black arrows in the figure represent the transfer of these results. It can also be seen that there is interconnectivity between the different themes since results from theme 6 (*Social integration and governance* are used as an input for the research in theme 2 and the results of theme 2 are used as an input for theme 1 (*Heat networks and system integration*. This thesis is contributing to project 2C which focuses on: *Current and future cost of MT and LT district heating networks*. The goal of research 2C is to generate uniform cost parameters based on current construction methods and new construction methods that are being researched in research topic *2B. Smart construction methods*. As can be seen in the figure project 2C uses inputs from all other research topics in theme 2. However, since the main contribution of this thesis is finding correlations between cost and surrounding parameters considering similar infrastructures, this thesis is not dependent on the input from the other research topics in this theme. However, this thesis is dependent on other research that is being conducted in project 2C itself and this thesis could be used as a starting point for further research in project 2C.

There are a lot of different companies involved in the research consortium. However not all the involved companies are also directly involved with this thesis since different companies contribute to different topics. The companies directly involved in topic 2C and therefore informed and asked for inputs regarding this thesis are: *Heijmans*, *Enpuls*, *Firan*, *Eneco*, *Ennatuurlijk*, *Deltares* and *TU Delft*.

1.7. Structure of report

The rest of this thesis is structured as follows: First, in chapter 2 a state of the art analysis on cost calculations tools for DH networks is conducted to identify opportunities to increase the prediction accuracy of DH construction costs. Both the currently available energy transition models that are used in professional practice and relevant research from the scientific community is discussed. In this chapter subquestions, one and two are answered. Then, the proposed research methodology, which should lead to more accurate construction cost predictions, is described in chapter 3. In this chapter also some information on the chosen model design is presented. Afterwards, the interviews that were conducted with experts and the process of collecting project data is discussed in chapter 4. In the interviews with experts both the similar enough infrastructures and the surrounding parameters that are considered in this research should be identified answering subquestions three and four. The tuning and training of the statistical models as well as a dummy analysis using generated dummy data, that can be used to validate the modeling approach, are presented in chapter 5. In chapter 6 the resulting models and their predictive performance is discussed. The answers to the last two subquestions and the main research question can be arrived from this chapter. Finally, in chapter 7 and chapter 8 the discussion, conclusion and recommendations can be found.

2
State of the Art Cost Calculations

In this chapter, a state of the art overview of cost calculations for DH and similar infrastructures is presented. The chapter is divided into two sections. First in section 2.1 all currently existing energy transition models that calculate construction costs of DH networks are discussed. This first section gives an overview of the models used in professional practice. The second part of the chapter, section 2.2, is focused on the scientific approaches used for cost modeling in existing literature. This section particularly focuses on Machine Learning (ML) approaches for cost prediction. First, the absolute basics of ML are explained that are required to understand the literature that is presented in the rest of this subsection. Secondly, papers comparing different ML approaches are discussed. Finally, some papers that used ML for cost prediction of DH networks and similar infrastructures are presented.

This chapter has two main goals. First, it is important to identify how (energy transition) models used in professional practice calculate construction costs of DH networks. To validate that indeed, currently, models do not (sufficiently) include surrounding parameters in their cost calculations. Furthermore it is also identified how the generated construction cost model in this study can best be linked to the already existing models. Because this study, only focuses on the construction cost, whereas other energy transition models already have other cost components, like energy production costs, implemented. This means that it is desirable to merge the model from this study with the already existing models instead of developing an entirely new energy transition model. The second goal of this chapter is to identify the most promising cost modeling approach.

2.1. Current cost calculating models for District Heating

As discussed in section 1.1 an accurate estimation of the CAPEX of a DH networks is important in the design process and for choosing the "best" sustainable heating solution for a certain neighborhood. According to the "klimaatakkoord" [6] all municipalities should write a transition vision heat by 2021. To help climate policymakers at the municipalities all around the country with choosing the right sustainable heating solutions various energy transition models have been developed. These models calculate, among other things, the construction costs of DH networks. Most models, however, have a broader scope than calculating solely the costs associated with DH. But there are also energy transition models that don't calculate the CAPEX of HT at all and solely focus on for example electricity. To get an overview of the state of the art of (construction) cost calculating tools for district heating networks these energy transition models are analyzed. Since not all the energy transition models are interesting for this study, a selection is made based on the following criteria:

- The model calculates construction costs for DH networks
- The model is currently able to make calculations (out of the concept phase)
- A detailed description and/or the model itself is either open source available or the model is owned/developed by a company or organization that is willing to contribute to this research.

The following approaches are used to find relevant models for this study:

- Studying the "energietransitie rekenmodelen" overview pdf drafted by Netbeheer Nederland [47]

- Interviews with experts
- Literature study [29][52][21][32][26][49]

The models that were found and matched the above mentioned criteria are presented in Table 2.1 and Table 2.2. In the tables, important information about the models is presented. In subsection 2.1.1 four different categories for CAPEX calculation methods (mentioned in the tables) are defined and discussed, together with more elaborate explanations of some models applying the methods.

2.1.1. Different methods to Calculate CAPEX

The models shown in Table 2.1 and Table 2.2 use different formulas and/or parameters for their calculations. However, some of the models do adopt very similar methods to calculate construction costs. To get a better understanding of the similarities between these models they are subdivided into four categories. These categories are created in this research and divide models based on how they calculate construction costs. The four categories are described more elaborately in the following subsections. The main difference is the level of detail in which they calculate costs. Logically, methods with a higher level of detail consider more variables that could influence the costs. Where the first category has the lowest level of detail and category four has the highest detail level. It is, however, important to keep in mind that a model is only as good as its input data also known as the Garbage in Garbage out (GIGO) principle. Meaning that the model considering the most variables does not necessarily give you the most reliable result. Information on the input parameters and the validation strategies used in the models is provided in subsection 2.1.2.

Category 1: Number of Households

The first method for calculating CAPEX uses the number of households connected to a new DH network as the main indicator for the CAPEX. In this approach, all the costs (labor, pipes costs, HTS, etc) are considered but they are included in a single cost parameter. In the formulas in this report, all cost parameters are represented by a capital C followed by a small letter to define the unit of the cost parameters. For example, (Cp, Cm, Ch) are cost per power (kW), cost per meter, and cost per household respectively. The **Energie Transitie Model (ETM)** and **Energie Transitie Atlas (ETA)** use this approach to calculate costs. For calculating the construction cost of distribution pipes $(CAPEX_{distribution})$ the ETM model has a fixed price per meter pipe $(Cm_{distribution})$ and an average connection length per house $(\overline{L_{connection}})$ resulting in a average price per house. The average connection length per house is dependant on the percentage of houses connected to the network [1]. The construction costs of the transport pipes $(CAPEX_{transport})$ and the heat stations $(CAPEX_{stations})$ are proportional to the peak demand (P_{peak}) which in turn is proportional to the number of houses connected. The formulas used to calculate cost in the ETM are given in Equation 2.1.

The ETA model uses different connection costs for different types of houses and construction years. Also, every house type has a cost for the rural environment and one for the urban environment. Meaning that the construction cost in the ETA is dependent on the house type, building year, and neighborhood type (rural/urban). Because the model use different costs for different surrounding parameters the ETA could also be placed in the more detailed category three. In this category, surrounding types are included in the cost models. However, since the model does not use any dimensions or network characteristics like diameter, temperature, material type, etc, which is a requirement to be in category 2 or higher the model is placed in category 1. The formulas used to calculate cost in the ETA model are shown in Equation 2.2.

ETM[1]:

$$CAPEX = CAPEX_{transport} + CAPEX_{distribution} + CAPEX_{stations} \tag{2.1a}$$

$$CAPEX_{transport} = Cp_{transport} * P_{peak} \tag{2.1b}$$

$$CAPEX_{distribution} = Cm_{distribution} * \overline{L_{connection}} * N_{households} \tag{2.1c}$$

$$CAPEX_{stations} = (Cp_{HTS} + Cp_{sub}) * P_{peak} \tag{2.1d}$$

ETA[48]:

$$CAPEX = Ch_{rural/urban} * N_{households} \tag{2.2a}$$

$$Ch_{rural/urban} = f(B_{type}, B_{year}) \tag{2.2b}$$

Table 2.1: State of the art cost calculating models

Name Model	Owner / open source	End user(s)	Category	Sector(s)	Scale of model	Remarks
VESTA mais	PBL / open source	Municipalities and consultants (for municipalities)	2	Heating sector, electricity is considered but just for fulfilling heat demand	Netherlands	Supported by dutch government, every municipality gets report with results
Energie-transitie-model	Quintel / open source	Policy makers	1	All energy	Netherlands	Results are better on bigger scale, model is very user friendly.
Thermos	Consortium managed by centre for sustainable energy / Open source	Local decision makers	3/4	District heating	Everywhere, in Europe one district heating network location can be chosen (size specified by hand). Some areas have more standard data available	Funded by EU horizon 2020 project
Planheat	Open source	Consultants for government parties (also municipalities, and provinces)	2 / 3	Heating sector	Everywhere in Europe a city or district can be chosen as simulation scope	Funded by EU horizon 2020 project
Caldomus	Innoforte	Municipalities and housing cooperation's	2	Heating sector	Neighborhoods in the Netherlands	Model is quite similar to Vesta (regarding heat net costs calculations)
RETscreen	Government Canada / open source	Decision makers and energy experts	3	All energy	Any size up to house scale	Model is downloaded over 200.000 times worldwide
Comsof Heat	Comsof	Urban planners	2	District heating	Choose own area, software will automatically generate best network in this area. Possible to exclude certain houses or streets from possibilities.	For calculating costs model is dependent on hand picked input data from similar projects. Possible to make certain streats or areas more expensive to build.

Table 2.2: Overview state of the art energy transition models part 2

Name Model	**Owner / open source**	**End user(s)**	**Category**	**Sector(s)**	**Scale of model**	**Remarks**
Heat	Alliander (omons) / private	Regional governments, housing cooperations and network operators	2	District heating	Neigbourhoud up to municipality scale.	Model is used less since Alliander decided to cut funding for Omons their daugther company who where responsible for the model.
DEEB	Eneco / private	Employees Eneco	3	Heating sector	Single building up to entire neigbourhoud	Model is only availeble for employees of Eneco but can be used together with municipalities or other parties when they work together with Eneco
Startmotor-kader	ECW (expertise centrum warmte) / public	Housing cooperations and heating companies	2	District heating	Size of one district heating network which is dependent on seperatate network design outside of the model. The model solely uses pipe lengths for calculating cost not area's.	Tool constructs business case, design of the network and corresponding cost should be determined by hand or with different tools. Standard cost parameters from VESTA are included for first indication calculation.
Energie transitie atlas	Overmorgen (arcadis) / private	Local and regional Governments	1[1]	Heating sector	Neigbourhoud	(shielded) online GIS viewer presents results. Possible to get "raw"data output as well
Chess	TNO / private	Energy companies and network operators	2	District heating	One DH network	Focussed on phycics of DH networks not on costs
CEGOIA	CEDelft / private	Policy makers from local and regional governments	2	Heating sector	Neigbourhood up to entire country (Netherlands)	Results can be presented in GIS based images

Category 2: Dimensions and characteristic network

The second method for calculating CAPEX takes a more in-depth approach and considers the dimensions and characteristics of the DH network. The temperature of the network, the length of the pipes, and the heat demand that has to be full-filled are typical parameters that are considered. Taking this approach, different pipe diameters can be considered and a trade-off between a higher temperature or thicker pipes to full-fill the same demand arises. However, most of the models take temperature as a given input and adjust the diameters and corresponding costs. Examples of models using this method are **VESTA** and **Caldomus**. According to interviews with market parties conducted by PBL [54], when validating their VESTA model, the temperature levels have a relatively low impact on the pipe costs compared to the required demand and the length of the pipes. That is why in the latest version of the VESTA model the costs per meter pipe Cm_{pipe} are solely based on demand. Different approaches are used for different network types. For transport pipes, the distance from the heat source to the neighborhood (L_{source}) multiplied with a detour factor (f_{detour}) is used to calculate cost and the average distance a house is from the road in a certain neighborhood ($\overline{L_{dis_{r}oad}}$) is used for connection pipes. For distribution pipes, different approaches are used based on the temperature [54]. In Vesta also management cost ($Cost_{man}$), compensation cost ($Cost_{com}$) and cost for HTS and heat buffers are considered . The formulas used in VESTA to calculate cost are shown in Equation 2.3. In the Caldomus model, the construction costs for transport pipes are dependant on length ($L_{trans\_network}$), temperature difference (ΔT) and peak demand [37] and the construction cost for distribution pipes are based on a price per meter and a relation which calculates the length of the pipes based on the total network area. The formulas used in the Caldomus model are shown in Equation 2.4.

VESTA[54]:

$$CAPEX = Cost_{man} + Cost_{com} + CAPEX_{transport} + CAPEX_{distribution} \tag{2.3a}$$

$$+ \sum_{connection=1}^{N_{households}} CAPEX_{connection} + CAPEX_{stations} + CAPEX_{buffer} \tag{2.3b}$$

$$CAPEX_{transport} = Cm_{pipe}(P_{peak}) * L_{source} * f_{detour} \tag{2.3c}$$

$$CAPEX_{distribution} \subset \begin{cases} WKO\&TEO = (600 * \frac{187}{15741}) * A_{network} \\ LT = Cm_{pipe}(P_{peak}) * f_{LT} * \sqrt{A_{network}} \\ MT = Cm_{pipe}(P_{peak}) * f_{MT} * L_{road} \end{cases} \tag{2.3d}$$

$$CAPEX_{connection} = \overline{L_{dis\_road}} * Cm_{pipe}(P_{connection}) * N_{connections} \tag{2.3e}$$

$$CAPEX_{stations} = (Cp_{HTS} + Cp_{sub}) * P_{peak} \tag{2.3f}$$

$$CAPEX_{buffer} = Ch_{buffer} * N_{households} \tag{2.3g}$$

$$Cm_{pipe}(P) \subset \begin{cases} min = 400 + 210 * (P * 0.001)^{0.5} \\ max = 800 + 200 * (P * 0.001)^{0.6} \end{cases} \tag{2.3h}$$

Caldomus[37]:

$$CAPEX = CAPEX_{transport} + CAPEX_{distribution} + CAPEX_{stations} \tag{2.4a}$$

$$CAPEX_{transport} = L_{trans\_network} * (\Delta T)^{-0.189} * 624.62 * (P_{peak})^{0.1893} \tag{2.4b}$$

$$CAPEX_{distribution} = Cm_{distribution} * L_{dis\_network} \tag{2.4c}$$

$$L_{dis\_network} \subset \begin{cases} primary_{dis} = f_{pri} * 2 * \sqrt{2} * \sqrt{A_{network}} \\ secondary_{dis} = f_{sec} * N_{sub\_station} * 0.25 * 0.5 * \sqrt{2} * \sqrt{A_{network}} \end{cases} \tag{2.4d}$$

$$CAPEX_{stations} = (Cp_{HTS} + Cp_{sub}) * P_{peak} \tag{2.4e}$$

Category 3: Surrounding type

The third method for calculating CAPEX is very similar to the second method. However, there is one important difference in this third category also the surrounding type is considered in calculating costs. Resulting in the fact that a meter of pipe in a highly urban area is more expensive than that same meter of pipe in a

[1] ETA Model uses surrounding category values (building year, house type, neighborhood type) for cost calculations but no network dimensions or characteristics are included

rural area. However, it is important to realize that on average shorter pipe lengths are required in urban areas reducing costs compared to rural DH networks. This method is applied by for example the **RETscreen** model which makes a distinction between the two surrounding types urban and rural in the formulas shown in Equation 2.5 the surrounding type is S_{type}. As can be seen the exact formulas for the CAPEX calculations of the different network components are not available in the literature and, therefore, unknown in this research. However, by using the software and altering input parameters it is analyzed that certain parameters influence the different costs. The results from this analyses are given in Equation 2.5b, Equation 2.5c and Equation 2.5d. Where $(L, S_{type}, d, f, N_{stations})$ represent, the length of the pipe, the surrounding type, the diameter of the pipe, factors to alter cost (default 1), and the number of heat stations respectively.

RETscreen:

$$CAPEX = \sum_{i=1}^{N_{trans\_seg}} CAPEX_{transport} + \sum_{j=1}^{N_{dis\_seg}} CAPEX_{distribution} + \sum_{j=1}^{N_{stations}} CAPEX_{stations} \tag{2.5a}$$

$$CAPEX_{transport} = f(L_i, S_{type}, d_i, f_{trans}) \tag{2.5b}$$

$$CAPEX_{distirbution} = f(L_j, S_{type}, d_j, f_{dis}) \tag{2.5c}$$

$$CAPEX_{stations} = f(S_{type}, f_{stations}, N_{stations}) \tag{2.5d}$$

Category 4: Actual surrounding parameters

The final category again is similar to categories 2 and 3, however, in this category the surrounding is not taken into account by selecting a surrounding type but actual surrounding parameters like for example road types, traffic density, ground quality, type of buildings, etc. are considered when calculating costs. This final category might be the most detailed but is also the hardest to implement since a significant number of historical data is required to include such surrounding parameters in the calculation. In this research, no model was found which applies this method. The **Thermos** model, which uses the most detailed approach found in this research, applies a method that depends on the users' settings and inputs gets the closets to this "ideal" situation. In the Thermos model, the pipe costs are divided between mechanical engineering costs and civil engineering costs. The mechanical engineering cost include:

- Buying pipes
- Welding pipes together on site

And the Civil engineering cost consist of:

- Digging up road surface
- suspending traffic
- suspending parking

For every pipe segment considered in the calculation, these costs are calculated using Equation 2.6b and Equation 2.6c respectively. As can be seen in the equations the cost is, besides length and diameter, also dependant on the parameters f_{mech_1}, f_{mech_2}, f_{civ_1} and f_{civ_2}. These parameters can be set for every pipe segment or default parameters can be chosen for certain pipe types (f_{mech_1} and f_{mech_2}) or certain roads and or surrounding types (f_{civ_1} and f_{civ_2}). Default values for the civil parameters provided on the Thermos help page are shown in Table 2.3. It is important to note that these values are based on historical data from projects in the UK. When using the thermos tool for calculations outside of the UK these values should be re-calibrated using historical cost data from local projects. This means that to include local restrictions in the cost calculations outside of the UK these four parameters should be generated based on historical cost data of similar surroundings.

Thermos[5]:

$$CAPEX = \sum_{i=1}^{N_{pipe\_segments}} Mechanical_{cost}(i) + \sum_{i=1}^{N_{pipe\_segments}} Civil_{cost}(i) \tag{2.6a}$$

$$Mechanical_{cost}(i) = L_i * (f_{mech\_1}(i) + (f_{mech\_2}(i) * d_i)^{1.3}) \tag{2.6b}$$

$$Civil_{cost}(i) = L_i * (f_{civ\_1}(i) + (f_{civ\_2}(i) * d_i)^{1.1}) \tag{2.6c}$$

Table 2.3: Sensible civil parameters for the UK

Location	Surface	f_{civ_1}	f_{civ_2}
Urban	Hard	1200	500
Urban	Soft	450	0
Suburban	Hard	850	200
Suburban	Soft	100	0

2.1.2. Validation of the models

As discussed in subsection 2.1.1, not just the level of detail of the model is relevant for the accuracy of the generated results. Also, the used parameters and the validation strategy of a model are important factors that influence the accuracy. In Table 2.4, for every model a short description of the validation strategy and used input parameters is given. In the last column, it is also stated whether or not it is possible to use own cost parameters as an input for the model calculations. As can be seen in the table most models offer the possibility to enter cost parameters yourself and for some models, it is even mandatory. There are also models, Thermos and RETscreen, which do have validated cost databases but only for the countries in which they were developed, UK and Canada. Which means that using these cost parameters in the Netherlands might not give results that are as accurate as they are in the UK and Canada and should, therefore, be used with caution.

2.1.3. Room for improvement

Looking at how the studied models consider surrounding factors in their construction cost calculations the conclusion can be drawn that there is room for improvement. As can be seen in Figure 2.2 none of the researched models takes actual surrounding parameters into account, and most models only consider dimensions (diameter, length) and characteristics (temperature, network type) for their construction cost calculations. Not considering surrounding parameters in construction cost predictions is a problem because the uncertainty of the predictions is higher and prediction will be less accurate when the surrounding parameters do influence the cost, which is very likely. When for example considering the VESTA model, which is a leading model in the Dutch energy transition sector, it can be seen that the bandwidth for cost calculations is very wide. Comparing the two lines in Figure 2.1, which depict the maximum (green) and minimum (blue) pipe price per meter of the VESTA model, it can be seen that the upper bound is approximately twice as high as the lower bound for the depicted power range and this difference will only grow for higher powers. In the VESTA documentation, it is stated that this difference is mainly caused by the difference in road types (closed roads are more expensive than open fields or clinker pavers). But other factors like crammed subsurface or ground pollution can cause serious price fluctuations as well [54]. Researching the impact of surrounding factors on the construction cost of DH networks is likely to reduce this bandwidth and, therefore, increase the accuracy of the energy transition models.

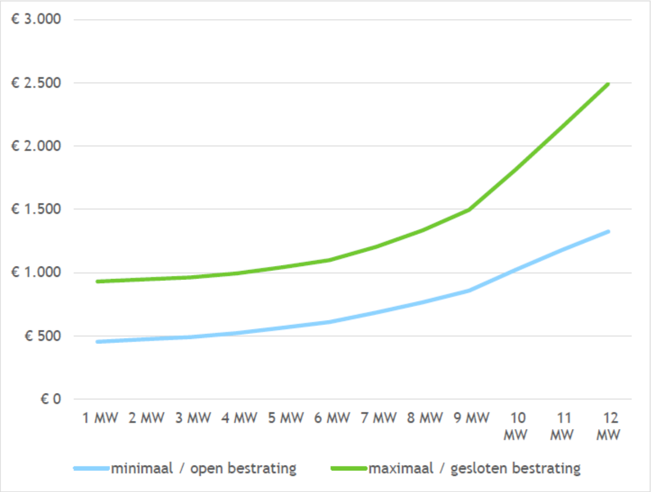


Figure 2.1: Construction costs per meter pipe used VESTA model [54]

Especially since, a lot of models offer the opportunity to input your own cost parameters and some models even offer the possibility to do that for specific pipe segments, roads, or areas. When looking at Figure 2.2 again, all models with underlined names offer the possibility to enter your own cost parameters. This means that surrounding-based cost parameters can be implemented in a lot of existing models and potentially improve the accuracy of these models significantly. Meaning that without making an entirely new energy transition model, which should also include other important costs like maintenance and energy production costs, a surrounding based construction cost model will be of added value to the energy transition. Moreover, such a construction cost model could potentially also be used in combination with models not found in this study, or it can be used separately without a supporting model to identify areas in the Netherlands where constructing a DH network is relatively cheap.

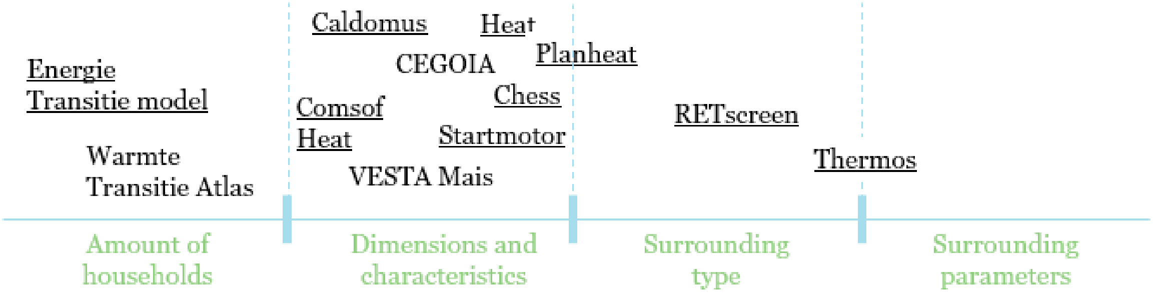


Figure 2.2: Overview of energy transition models, divided based on how they include calculate construction costs

Table 2.4: How models gather their input parameters and whether they applied a validation strategy

Model	Description input parameters and validation strategy	Input own cost parameters
Vesta Mais	The parameters used in the Vesta Mais model were validated in multiple validation sessions with relevant market parties. In these sessions, the used parameters were shown to companies, and feedback was received about whether these (cost) parameters were inline with current market prices.	Not possible (In theory it is possible but in practice it requires a lot of knowledge about the model to change parameters)
ETM	No strategy to validate the parameters used for calculating district heating cost is found in the extensive online available documentation. However, it is stated that the model parameters for district heating are based on the Vesta Mais model.	Possible (as percentage difference compared to given base case)
Thermos	For the UK some suggestions for parameters are given, but it is recommended that the user defines his own cost parameters based on available historical data.	Possible and recommended
Planheat	It is possible to use own cost parameters as well as standard parameters included in the model. Unknown whether these standard parameters are validated in the market. The outcome of the entire model is validated in multiple cities and found to be trustworthy (models does more than calculating construction costs).	Possible
Innoforte	In Innofortes documentation about the model, it is stated that the used cost parameters are based on literature and historical projects. It is, however, unknown whether these parameters are validated or are being updated regularly.	Not possible
Retscreen	Possible to enter own data or use cost database build into the Retscreen software, possible to give weight factors to certain cost components to make them relatively more or less expensive. The weight factors can be used to compensate for local restrictions or opportunities. Cost data validated for Canada.	Possible and recommended especially outside of Canada.
Comsof heat	Cost data should be entered by hand based on similar projects.	Mandatory
Heat	Cost data is based on experience and historical data that was available to Qirion and its partner companies (Alliander) Unknown if model is validated in the market.	Possible
DEEB	Cost data is based on historical cost data available to Eneco. It is estimated that the bandwidth of the cost calculations is approximately 20%.	Not possible
Startmotor-kader	Default cost data is available (mostly based on VESTA) but it is recommended to input own cost data taking into account local restrictions and opportunities.	Possible and recommended
Warmte transitie atlas	Cost are “based on experiences in the market”, In the documentation received about the WTA nothing was mentioned about the validation of this data or the results.	Not possible
Chess	In the so-called post-process cost are determined together with clients. The physical properties which are designed by the model are linked with costs based on experience and knowledge. It is important to note that the cost is currently not an important part of the model but should be seen as a bonus feature.	Mandatory
CEGOIA	The parameters used in the CEGOIA model are probably based on VESTA. It is important to note that there is no publicly available statement about this. But CEdelft, the owner of CEGOIA, is directly involved with the development of the VESTA Mais model.	Not possible

2.2. Cost modeling literature review

Estimating the (construction) cost of different kinds of projects can be done in a lot of different ways. These cost predictions can be based on historical data, knowledge of experts, or a combination of both of these. Also, the approach to predicting cost can be different. The different approaches for modeling cost can be roughly categorized in following three categories [36]:

- Analogous cost estimation
- Parametric cost estimation
- Bottom-up cost estimation

The basis of **analogous cost estimation** is predicting the cost of a project/product based on similar projects/products and the relative differences between them. The underlying assumption here is that similar projects have similar costs. The downside of this approach is that it is impossible to predict the cost of projects which are not similar "enough" to already existing projects. Also experts' knowledge is required to identify which project properties, and thus cost, are similar and which are not [36].

Parametric cost estimation is about building so-called "Cost estimation relations". This basically means finding mathematical relations ships between project characteristics (like size, weight, and material) and construction cost. The downside of this approach is that the found mathematical relations are only validated for a certain range (available historical data). This for example means that, although it is possible to predict the cost of a project which is a lot larger than any historically realized project in the database the result is likely to be inaccurate[36].

The final and most detailed cost prediction method, **bottom-up cost estimation**, includes costs estimations for all steps of the construction process including for example materials, labor, and transport cost. This cost estimation method results in a very detailed cost prediction for which no historical data is required. The downside of this approach is that a lot of in-depth knowledge and experience with the construction process is required to properly set up the cost model. It should also be noted that besides the three modeling approaches discussed above a so-called rule of thumb or expert judgment can also provide useful cost estimations in the early design stage[36].

For the estimation of construction costs of DH networks in thesis, a combination of the first two cost estimation approaches is applied. The third approach is not considered for this research because of two reasons. First, the required experience in constructing DH networks in practice isn't available at both Deltares and the TUDelft. Second, the detail level resulting from a bottom-up-approach is not needed and the required inputs are not available in the early design phases for which the energy transition models discussed in section 2.1 are used. The results of this thesis are meant as a contribution to this early design phase.

For the cost predictions in this research first the **Parametric cost estimation** approach is used, to find mathematical relations to estimate the construction cost of natural gas and drinking water networks. Secondly the **analogous cost estimation** approach applied. In this seconds step it is explained how the cost estimation models for similar infrastructures can be used as inputs for a model that predicts DH construction cost. Since, as explained more elaborately in chapter 3, unfortunately, there is no data available on DH projects this seconds step cannot be fully finished. This is because it is important to also have heat data to validate that indeed the water and gas models are similar enough to heat. Without this heat data, only speculations can be made about the construction cost of heat networks based on the known similarities and differences between the considered similar infrastructures and heat networks. These similarities and differences are studied more elaborately in section 4.1.

Finding the mathematical relations to estimate the construction cost of fluid infrastructures can most easily be realized by using ML. However, there are a lot of different machine learning algorithms available. To identify the best suited ML method for this thesis a literature review is conducted. Characteristics of different ML algorithms are studied and compared and similar research projects were studied to identify which ML methods were applied in these research projects.

The rest of this section summarizes this literature review presenting an overview of existing literature on cost modeling using ML. The literature is subdivided in three categories. First some general background information on ML is provided in subsection 2.2.1. Some papers comparing different ML approaches for cost modeling are discussed in subsection 2.2.3. Then existing literature on cost modeling of DH networks is presented in subsection 2.2.4. Finally, in subsection 2.2.5 an overview of literature on cost calculations of similar infrastructures is discussed.

2.2.1. Different machine learning types

In this subsection, some very basic knowledge on ML is presented that is required to understand the literature presented in the next subsections. When one already possesses this basic knowledge on ML this subsection can be skipped and reading can be continued in subsection 2.2.3. More information on the specific ML aspects that are applied in the construction cost model, developed in this research, can be found in chapter 5.

In the field of ML two different types of "learning" can be distinguished, Supervised and unsupervised ML. In unsupervised ML the goal is to find unknown patterns in existing unlabeled data sets. As the name would suggest this process is without "supervision" (inputs) from humans meaning that the outcome of such a model cannot be predicted. In other words, you cannot tell for such a model which relation should be discovered. This kind of ML is not applied in this thesis because the goal of this research is to find a specific relation (relation between costs and other inputs) in a data set rather than an unknown random relation. In supervised ML input data (**x**), also known as the independent variables, is labeled with a corresponding output (**y**), also known as the dependent variable. The label (y) of the data point in essence is the state that has to be predicted by the model in the future. In cost prediction, the label (**y**) would be the construction cost and the input (**x**) would be project characteristics like material type, dimensions, ground pollution level, etc. The goal of supervised machine learning is to train a model using labeled input data (x_{train}). This model can be used to predict a new unknown output (y_{new}) corresponding to a new input state (x_{new}).

Input data (**x**) to any ML algorithm can be either numerical or categorical, for some examples of this see Table 2.5. Furthermore, a machine learning algorithm can also take both numerical values and categorical values as inputs for the same model. The only difference is that a categorical input which can have a lot of different forms is not ideal since all of them will be considered separately, meaning that the model will become more complex compared to a numerical input which can take in theory an infinite number of different values without getting more complex. The reasoning behind this is different for different algorithms but for all of the algorithms categorical values with a lot of different potential categories are undesirable.

Table 2.5: Examples of differences numerical and categorical inputs of Machine Learning models [42]

Variable type	**Example**	**Handled as**
Continuous number	$3.5[m], 14.24[m], 6.0[m]$	Numerical
Discrete number with ordering	0 houses, 1 house, 2 houses	Numerical
Discrete number no ordering	1 = Rotterdam, 2 = Amsterdam, 3 = Utrecht	Categorical
Text string	Local road, Street , Other	Categorical

Besides input data, the output data (**y**) of a ML model can also be both numerical and categorical. The difference compared to input data is data is that a certain machine learning model can either provide a numerical output or categorical output and both of these options have a different generic name. When the output data of a model is numerical the model is called a **Regression** model and when the output data is categorical the model is called a **Classification** model. These two different types of ML models have different properties and require different methods for training (solvers).

Table 2.6: Machine learning categories with examples

Name	**Example**	**Classification / Regression**
Distance based	k-nearest neighbors (k-NN)	Both
Rule based (decision tree)	Classification tree	Classification
	Regression tree	Regression
	Random forests	Both
	Gradient boosted trees (GBS)	Both
Parametric based	(Multivariate) linear regression (MLR)	Regression
	Logistic regression	Classification
Black box	Generalized additive models (partly parametric)	Both
	Neural networks (NN)	Both
	Support vector machine (SVM)	Both

Besides the distinguishing models based on their output ML models can also be subdivided based on their working principle. Four different categories of ML models are given in Table 2.6 together with some examples [38]. As can be seen in the table most ML approaches can be both used for classification and regression. For example, classification trees and regression trees have the same working principle and are also known as decision trees. However, when used for regression a decision tree is called a regression tree and when used for classification it is called a classification tree. The same holds for linear regression and logistic regression who also have the same working principle. However, the name logistic regression is a bit more tricky since this machine learning method is a classification model when the name would suggest otherwise.

k-Nearest Neighbors:

The first method in the table, **k-NN**, is founded on the assumption that the output y_{new} of a new unknown state with input x_{new} is likely to be the most similar to the output y_{train} of the training data points of which the input x_{train} is the most similar to x_{new}. To implement this, the distance between the new input x_{new} and all the training data points x_{train} is computed [1]. A common approach to calculate this distance is the so-called Euclidean distance, this distance (D_{euc}) can be calculated using Equation 2.7 [60]. Where n is the number of input variables in each data point ($\mathbf{x}$). The k (integer user input) historical data points which are the closest to the new input x_{new} all have a vote for the prediction of y_{new}. For regression, this means that the average of the k closest data points is the prediction of y_{new} and for classification, a winner takes all strategy is applied.

$$D_{euc} = \sqrt{\sum_{i=1}^{n} \left(x_{new}^{i} - x_{train}^{i}\right)^{2}} \tag{2.7}$$

Decision Trees:

The rule-based method also known as a **decision trees** are exactly what the name would suggest. A binary (decision) tree is used for making predictions. A decision tree divides the input space (all potential x_{new}) into smaller regions according to logical (binary) rules. The three begins in a so-called *root node* where all new data "enters". Starting from this root note a data point travels down the tree through *internal nodes* which are connected by *branches*, every internal node has its own binary splitting rule (continue left or continue right). The data point follows this route until it reaches the end of the three in a so-called *leaf node*. All the leaf nodes of a decision tree have a unique prediction of y_{new} which is based on the training data. The prediction in a leaf node follows the same logic as the k-NN algorithm. However, instead of using the "closest" k data points (k-NN) all the data points from the training set that ended in the same leaf node are used to make the prediction (average for regression, winner takes all for classification). Depending on the chosen settings for training a decision tree it is possible that a leaf node only contains one data-point meaning that the prediction is identical to that data point.

The shape of a decision tree is dependent on the training data and the user settings for training a regression tree. Some examples of common training settings are maximum tree depth and the minimum amount of data points in a leaf node. If these limitations are not implemented the resulting decision tree will have only one data point in every leaf node. Which is very likely to result in so-called **overfitting** of your training data. Meaning that your model can perfectly predict your training data set but is unluckily to have good predictions for new unseen data. Overfitting is a common problem in ML that can happen with all machine learning models. For an visual example of a very simple (classification) decision tree using only two input variables (x_1, x_2) see Figure 2.3 [38].

[1] Distance between data points can be calculated in different ways and is also possible for categorical values

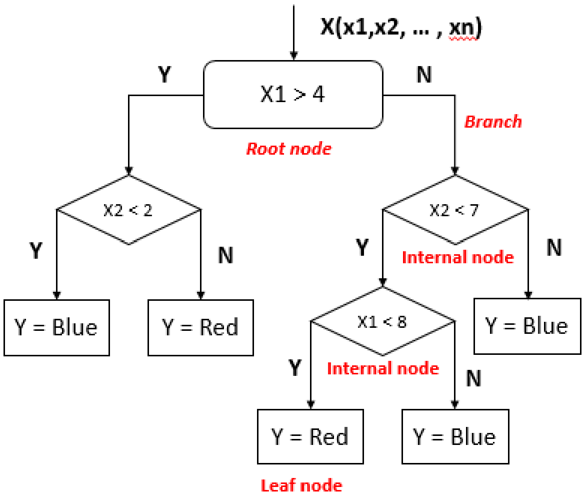


Figure 2.3: Example decision tree

Linear and Logistic Regression:

Linear regression is probably the most simple ML algorithm out there. A linear regression model is a parametric function, see Equation 2.8, of which the parameters ($\boldsymbol{\theta}^{\top}$) are estimated by minimizing a so-called cost function using data from the training database (training the model). A common example of such a cost function is the squared error loss see equation Equation 2.9[38].

$$y_{new} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \varepsilon = \begin{bmatrix} \theta_0 & \theta_1 & \dots & \theta_p \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_p \end{bmatrix} + \varepsilon = \boldsymbol{\theta}^{\top}\mathbf{x} + \varepsilon \tag{2.8}$$

$$L(y_{new}(\mathbf{x};\boldsymbol{\theta}), y_{train}) = (y_{new}(\mathbf{x};\boldsymbol{\theta}) - y_{train})^2 \tag{2.9}$$

$$y = \theta_0 + \theta_1 x_1^2 + \theta_2 ln(x_2) + \theta_3 x_1 x_2 = \theta_0 + \theta_1 x_i + \theta_2 x_j + \theta_3 x_k \tag{2.10}$$

$$\mathbf{x} = \begin{bmatrix} x_i & x_j & x_k \end{bmatrix} = \begin{bmatrix} x_1^2 & ln(x_2) & x_1 x_2 \end{bmatrix} \tag{2.11}$$

Maybe the most challenging about linear regression is the different names that are used for very similar models. First of all as already stated above Linear regression is a regression model, however, logistic regression is a classification model. Secondly, the word linear in linear regression also isn't that well-chosen since it is possible to use a non-linear regression function in linear regression. The word linear in this case means that a linear relation needs to be found between several non-linear terms. An example of this is shown in Equation 2.10 and Equation 2.11. Finally, the words multi, or multivariate are sometimes added to the name to indicate that the regression function takes multiple input variables.

Black box models:

Black box models like, **Neural networks**, **Support vector machines** and **Generalized additive models** are more complex than the models so far described above. These more complex models have the advantage that, when enough training data is available, they make more accurate predictions than less advantaged alternatives like (multiple) Linear regression[42]. The downside of these models, however, is that they are harder to train, require more computing power, and most importantly they are less interpretable [42]. The fact that the results of these models are not interpretable easily is a problem because the goal of this research is to compare the modeling results of different infrastructures. How these models work precisely will, therefore, not be discussed further in this thesis since it is outside of the scope. The results these models can achieve will be discussed in more detail below.

2.2.2. Scoring of a machine learning model

When choosing or fine-tuning a ML model it is very important to know how well a certain model is performing. This way it can be verified whether a certain change to a model improves the model's ability to predict a certain output. To score a model test data (y_{test}, x_{test}) is used. The x_{test} is inputted in the model to generate a model prediction y_{new}. This prediction can then be compared to true output value y_{test} to see how well the model can predict unseen test data. To measure how well a ML model is performing several scoring criteria exist. These scoring criteria are different for classification and regression models.

Scoring criteria of classification models:

Some examples of popular scoring criteria for classification algorithms are [38]: Accuracy, Recall, Precision, F_1-score. To understand how these four different scoring criteria work it is important to first take a look at the so-called confusion matrix shown in Table 2.7. In the confusion matrix, the four different scenarios for comparing y_{new} with y_{test} are presented and labeled. Using the labels from the table as inputs for the formulas presented in Equation 2.12 will result in the four scoring criteria mentioned above.

Table 2.7: Confusion matrix [38]

	$y_{test} = -1$	$y_{test} = 1$	total
$y_{new}(\mathbf{x}) = -1$	True neg (TN)	False neg (FN)	N*
$y_{new}(\mathbf{x}) = 1$	False pos (FP)	True pos (TP)	P*
total	N	P	n

$$Accuracy = \frac{TN + TP}{n} \tag{2.12a}$$

$$Recall = \frac{TP}{P} \tag{2.12b}$$

$$Precision = \frac{TP}{\mathrm{P}^*} \tag{2.12c}$$

$$F_1 - score = \frac{2TP}{\mathrm{P}^* + P} \tag{2.12d}$$

Scoring criteria of regression models:

For regression models also several scoring criteria exist. Some popular scoring criteria for regression models are [63]:

- F-test and t-test
- Coefficient of determination (R^2)
- Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE) and Normalized Root Mean Square Error (NRMSE)

The **F-test** is a statistical test, which calculates the change that a certain statistical model only found a correlation based on random chance, meaning that in fact, the model does not provide any useful insights. In theory, it could be true that the relations found by a certain model are only true for the used training data and not for the entire population. Since training data is (almost always) only a sample of the entire population this means that the results of the model are dependent on the random "selection" of this sample (training set). For example, in this thesis, not all networks that are ever constructed are considered only the ones that were provided by partners. If another training data set was used the model could have found entirely different relations or in the most extreme scenario no relation at all. This is most easily explained by considering the linear regression model shown in Equation 2.8. All the parameters ($\boldsymbol{\theta}^\top$) are estimated based on training data. When the training set is altered the predicted parameters ($\theta_0 = \theta_1 = \ldots = \theta_p$) will change. This means that the estimated thetas are actually a random distribution instead of a fixed number. If the standard error of this distribution is large enough this could mean that a found thetas that is positive for this training data set is zero or negative for another random sample of the data (Other projects are considered). An F-test test the change that all the parameters in ($\boldsymbol{\theta}^\top$) are equal to zero. If this null hypothesis ($H_0 : \theta_0 = \theta_1 = \ldots = \theta_p = 0$)

is true this would mean that the found model is absolutely useless. Logically, the lower the change that all the thetas are equal to zero the better the model performs. A **t-test** is very similar to an f-test. However, a t-test calculates the change of a single input parameter to be useless ($H_0 : \theta_0 = 0$). This test can, therefore, be used to check whether a certain input parameter delivers enough added value to the entire model. This test is very useful when tuning the model and choosing which potential inputs should be considered and which potential inputs should be ignored (not inputted anymore).

$$R^2 = 1 - \frac{\sum_i^n \left(y_i - \hat{y}_i\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2} \tag{2.13a}$$

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \mid \frac{y_i - \hat{y}_i}{y_i} \mid \tag{2.13b}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.13c}$$

$$NRMSE = \frac{1}{n} \sqrt{\sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \tag{2.13d}$$

All the next scoring criteria, $\boldsymbol{R^2}$ also known as the coefficient of determination, the **MAPE**, **MSE**, and **NRMSE** can be used in two different ways. First, these criteria can be used to estimate how good the model fits the training data (**Goodness-of-fit**). Secondly, the criteria can be used to validate whether a model can make accurate predictions (**Prediction accuracy**) for new and unseen data. The difference between the two approaches is whether training data (y_{train},x_{train}) or testing data is used (y_{test},x_{test}) to check the model performance. When training data is used the goodness-of-fit of a model is calculated and when testing data is used the prediction accuracy is estimated. The formulas that are used to calculate these above four criteria for both goodness-of-fit and prediction accuracy estimates are the same and are presented in Equation 2.13 [42]. In these formulas y_i is the actual dependent variables as provides in the data set (construction cost in this thesis), $\hat{y}_i$ is the predicted value, $bary$ is the average of all the dependent variables, n is the number of data points and k represents the number of inputs parameters.

Out of the above-mentioned scoring criteria, the $\boldsymbol{R^2}$ is the most commonly used to score ML models. An important remark about the above-mentioned scoring criteria is the fact that the value of R^2 should be as high as possible, the maximum for perfect model is 1, whereas for the other three scoring criteria lower values are desirable where 0 represents a perfect model.

Other ways to check model performance:

It is also possible to study the regression results by hand. There are three interesting things to look at when studying regression results. First, it is important that more or less the same number of data points are above the regression line as below, this is called homoscedasticity. When more than one input is used and the regression line cannot be plotted in 2D anymore the **Breusch Pagan** can also be used to check whether the residuals are homoscedastic. Homoscedasticity is important because it is one of the underlying assumptions of linear regression. So when the residuals are not homoscedastic the ML model is less likely to perform properly.

Another way of analyzing regression results by hand is by looking at the residual plot. In a residual plot one of the input parameters (x_p) is plotted against the error ($error = y_{new} - y_{train}$). If a machine learning model performs adequately all residuals plot should show random noise. If a certain pattern is distinguishable in a residual plot this means that this pattern is not found by the model and, therefore, the model could perform better than it is currently doing. When a pattern can be identified for all the data points that are predicted poorly, this pattern can be implemented in the model to improve the prediction for these extreme points. When this is successful the performance of the model is likely to increase. The downside of this approach is that it is a lot of handwork and it is hard to compare residual plots of different models.

A final way to potentially improve the model is by estimating whether the model overfits or underfits the training data. When the model **overfits** the training data the following things are likely to improve the performance [38]:

- Apply more regularization to reduce input variables (see subsection A.0.4 for short explanation of regularization)

- Apply bagging to decrease variance (see subsection A.0.2 for short explanation of bagging)
- Gather more input data
- Implement early stopping in the modeling settings

However, when the model **underfits** the training data the above mentioned things are not likely to help, but instead the following things should be applied [38]:

- Use a more complex model
- Use more input variables
- Use less regularization (see subsection A.0.4 for short explanation of regularization)
- Increase the training time

The question that remains is how can one estimate whether the model over or under fits the training data. A good strategy to get an indication for this is to calculate both the goodness-of-fit and the prediction-accuracy. When the goodness-of-fit is similar to or lower than the prediction-accuracy the model is likely to be **underfitting** the data [1]. However, when the goodness-of-fit is significantly higher than the prediction accuracy the model is probably **overfitting** the data.

2.2.3. Machine learning for cost modeling

In this literature review, a couple of papers are presented that compare the performance of different regression algorithms for cost predictions. Even though, it is possible to use classification for cost predictions, no papers that compared different classification algorithms were found since it is less common to use classification to predict the numerical value cost.

In research conducted in 2016 by Loyer et al. different ML methods were compared on their ability to predict the cost of jet engine components, in the early stages of the design process [42]. Five different ML learning approaches were chosen. Three new and relatively advanced ways of statistical modeling namely Generalized Additive Models (GAMs), Support Vector Regression (SVR), and Gradient Boosted Trees (GBTs), were compared to two more mainstream ML approaches (Neural Network (NN), linear regression). The dataset used for the case study (jet engine components) consists of 254 data points each containing 6 input variables (independent variables), the dataset is based on manufacturing cost in 2012. The performances of the 5 different ML algorithms are evaluated on 6 different criteria by the authors. In Table 2.8 the results of this analysis are presented. For the difference between Goodness-of-fit and Prediction accuracy see the explanation in subsection 2.2.1

Table 2.8: Comparison of the general characteristics of machine learning models Loyer et al. [42]

Model	Goodness -of-fit	Prediction accuracy	Interpretability	Easiness to fit and train	Extreme values	Computing affordability
MLR	+	+	+ + +	+ + +	+	+ + +
GAM	++	+	++	++	++	+
ANN	++	++	+	+	+ + +	+
SVR	+ + +	++	+	+	+ + +	++
GBT	+ + +	+ + +	++	++	++	++

The most important conclusions that can be drawn from this table are that linear regression models are the most interpretable and easy to train but are less accurate. When compared to the new and more advantaged ML models. The fact that regression models are less accurate can also be seen in Table 2.9 were 3 different scoring criteria for the 5 models are presented.

[1] goodness-of-fit is seldom lower than prediction accuracy since this will mean that the model performs better on unseen data than training data

Table 2.9: Comparison of the performance of the manufacturing cost models Loyer et al. [42]

Model	R^2	MAPE (%)	NRMSE (%)	Time (s)
Multiple linear Regression (MLR)	0.62	18.07	16.03	0.00
Generalized Additive Models (GAM)	0.82	13.05	12.19	0.23
Artificial Neural Networks (ANN)	0.88	11.30	11.02	0.30
Support Vector Regression (SVR)	0.93	9.15	9.66	0.17
Gradient Boosted Trees (GBT)	0.96	6.40	6.91	0.10

In another study, the performance of NN and linear regression is compared to predict the cost of building projects in the United States. The data consisted of 30 construction projects of continuing care retirement community buildings built in 14 different states from 1975 till 1995 [56]. The data set contained 7 independent input variables and the dependent variable: total project cost. For the linear regression model first, a selection was made on the most important input parameters. 5 out of the 7 original independent variables were included in the final model (**RM3**). In Table 2.10 it can be seen that the input variable with the highest p-value (t-test) is removed twice until only input variables with lower p-values remain. These 5 input variables are also inputted in two different NN models which differ in the number of hidden units in the model (more hidden units means more complex model). The first NN model (**NM1**) has 6 hidden units and the second model (**NM2**) has 3 hidden units.

Table 2.10: Selection process input variables regression model Sonmez et al. [56]

Model	Independent variables	R^2	Variable corresponding to the with the highest P value from t-test	P value of the coefficient coefficient
RM1	T, L, A, H, U, F, S	0.951	S	0.663
RM2	T, L, A, H, U, F	0.950	F	0.383
RM3	T, L, A, H, U	0.949	L	0.110

In Table 2.11 the Goodness-of-fit and prediction accuracy scores of the three ML models described above are presented. There are two important things to conclude from this table. First, when looking at the prediction accuracy it can be seen that linear regression outperforms both the NN models. This contradicts the conclusion drawn by Loyer et al. who concluded that NN have more accurate predictions than linear regression. This shows that different ML models perform better in different scenarios. In this specific case, linear regression likely outperforms NN because the training database is relatively small (only 30 data points). The second conclusion that can be drawn from the table, which is more straightforward, is that prediction accuracy scores are lower than goodness-of-fit scores since the prediction accuracy score is calculated based on unseen data.

Table 2.11: model performances Sonmez et al. [56]

Model	MSE (Goodnes-of-fit)	MAPE (Goodnes-of-fit)	MSE (Prediction accuracy)	MAPE (Prediction accuracy)
RM3	2.1×10^{12}	9.3 (%)	3.3×10^{12}	11.1 (%)
NM1	9.6×10^{11}	8.5 (%)	3.6×10^{12}	12.3 (%)
NM2	1.3×10^{12}	8.6 (%)	3.8×10^{12}	11.7 (%)

In yet another study a decision tree is compared to the k-NN method to estimate the construction cost index (CCI) in the United States. Data is collected for two relevant input parameters (CPI and PPI) and the CCI itself every month from Jan. 1985 till Dec. 2014 (348 data points) [60]. Both models are used to make predictions of the CCI on three different time scales namely: short-term, mid-term, and long-term. The results from this study are shown in Table 2.12. Three important conclusions from this table are: Modeling on the short-term (closer to your input data) is more accurate, Decision trees and k-NN have quite similar performances for this specific use case and the accuracy (MAPE) of all the models is a lot higher than for the previous two studies shown in Table 2.9 and Table 2.11 which indicates that some problems are a lot more suited for ML predictions than others. It is also important to note that this study had the largest database, which is probably a part of the reason why the models perform that well.

Table 2.12: model performances Wang and Ashuri [60]

Model	MAPE	MSE
Descision tree (short-term)	0.18%	501
k-NN (short-term)	0.19%	443
Descision tree (mid-term)	0.28%	1291
k-NN (mid-term)	0.35%	2146
Descision tree (long-term)	0.43%	2254
k-NN (long-term)	0.46%	2581

2.2.4. Machine learning for district heating

A leading researcher in the field of DH networks is Sven Werner. Werner, the (co)author of for example the publications about Heat Roadmap Europe and 4GDH both discussed in section 1.1, also published a paper together with Charlotte Reidhav about construction models for DH networks in areas with detached houses in Sweden [52]. In this publication also a literature review on existing studies is conducted. However, in this overview, only research from a couple of countries is presented (Denmark, Sweden, Iceland, Finland, and Germany). Three conclusions are drawn from this literature overview. First, the cost level of connecting detached houses to DH is equal in all the above-mentioned countries except Iceland. Because in Iceland "unique technical conditions lead to low cost"; Second comparing cost between different heating companies in Sweden turned out to be hard because different companies include different things in their construction costs; Finally the found cost estimation model available in Sweden only considers pipe length and number of houses when calculating construction cost. This means that variations in costs that are dependent on the surrounding of the project cannot be taken into account.

In the paper from Reidhav and Werner, a new approach is presented. Multivariate regression analysis is used to predict construction cost per house for DH networks, based on 55 projects between 1998 and 2005 in Göteborg Sweden. The following inputs were used for the initial multivariate regression analysis:

- The connection rate (connected houses divided by the total number of houses in the area)
- Pipe length per house
- Presence of ground frost
- Presence of tar asphalt in the asphalt layer
- Presence of rocks
- Pre investments due to future expansions

After fitting the model to the 55 projects the conclusion was drawn that out of the six inputs only the Pipe length per house (l_{house}) and the presence of tar in the asphalt layer ($boolean_{asphalt}$) had a high enough correlation with the project cost to be statistically significant. Resulting in the model described in Equation 2.14 which calculates the construction cost per house (Ch_{sweden}). The model has a R^2 value of 0.7 meaning that approximately 70% of the variations in construction costs are explained by the model.

Investment models for district heating in areas with detached houses[52]:

$$\boxed{Ch_{sweden} = 4230 + l_{house} * 232 + boolean_{asphalt} * 2360} \tag{2.14a}$$

In this research the conclusion was drawn that only the presence of asphalt and the Pipe length per house were significantly correlated with cost. This of course does not mean that other inputs, considered in this research or not, are per definition not correlated with the construction cost. When considering more projects and/or projects in different areas the results of the analysis can be very different.

Besides predicting cost per household as described above it is also possible to calculate construction cost per meter pipe that is installed. The former approach is, especially when also looking at similar infrastructures also see subsection 2.2.5 and the current (commercial) cost calculating models see section 2.1, more commonly used. Using a price per meter to calculate the construction cost of DH networks is for example used in a paper by Martin Leurent et al. where they do a *"Cost-benefit analysis of district heating systems using heat from nuclear plants in seven European countries"* [39]. In this paper two different formulas are used for calculating construction cost, one for distribution pipes ($Cm_{distribution}$) and one for transport pipes

($Cm_{transport}$) see Equation 2.15. In both formulas, the cost is only dependent on the diameter of the pipe(d_i). However, the cost function of transport pipes is non-linear whereas the cost function for distribution pipes is linear. This is because the two cost functions used in the research are based on different papers. The distribution function is based on a case study in Canada [28] which is also used for a study in Japan [57] and the transport function is based on a case study in Northern Poland [34]. The distribution cost function includes material costs, civil costs, sand filling, and labor costs and the transport function is based on a two-way buried pipeline with 200mm insulation and also includes costs for labor and pumping stations.

Cost-benefit analysis of district heating systems[39]:

$$Cm_{distribution} = 1570 * d_i + 235 \tag{2.15a}$$

$$Cm_{transport} = 3000 * d_i^2 + 4000 * d_i + 1500 \tag{2.15b}$$

2.2.5. Machine learning for similar infrastructures

The focus of this research is to use data of similar infrastructures to model DH costs. This is why also a literature research was conducted looking for cost modeling approaches for similar (pipe infrastructures).

Sewer systems:

In 2014 Valentino Marchionni et al. published a paper on cost modeling of sewer systems [45]. In this paper multivariate regression was used to construct cost functions of sewer systems using data from 17 projects in Portugal. The resulting cost functions for Gravity pipes are shown in Equation 2.16. As can be seen there are three different cost functions for three different materials ($Cm_{Concrete}$, Cm_{Iron}, Cm_{PPc}). All cost function calculate a price per meter and are only dependent on the diameter (d_i) and excavation depth (H_{trench}). It is also important to note that the coefficients of the three formulas differ very significantly with even the sign changing for some of them. When looking at for example the PPc function a very strong negative relation seems to exists between the diameter of the pipe and the construction cost per meter. This is contradicting common sense and most other cost models found in literature, so the formulas above should be used with care. It should be noted that only 17 projects were used to formulate these cost functions, which seems to be insufficient when using multivariate regression. A final important remark on the cost functions constructed by Marchionni et al. is that they used a so-called interaction term. **Interaction terms** are non linear input terms into a regression model by multiplying to different independent input variables, the terms H_{trench} and d_i in the regression model shown in Equation 2.16. These interaction terms are needed when the relation between an independent input variable (x_1) and the dependent variable ($\boldsymbol{y}$ is effected by the value of another independent input variable (x_2). This does not necessarily mean that x_1 and x_2 are correlated but it means that a certain parameter for example H_{trench} has a higher or lower impact on the dependent variable, Cm in this example, because another variable (d_i) is different. For this example, that means that the cost of digging a certain trench depth is not only dependent on the depth but also on the diameter of the pipe that needs to go into the trench.

Modelling Sewer Systems Costs with Multiple Linear Regression[45]:

$$Cm_{Concrete} = -203 + 125 * d_i + 131 * H_{trench} - 44 * H_{trench} * d_i \tag{2.16a}$$

$$Cm_{Iron} = 4.6 + 163 * d_i - 6.6 * H_{trench} + 147 * H_{trench} * d_i \tag{2.16b}$$

$$Cm_{PPc} = 384 - 1785 * d_i - 149 * H_{trench} + 794 * H_{trench} * d_i \tag{2.16c}$$

Besides the cost functions, a more valuable contribution of the paper by Marchionni et al. for this research is their overview of existing literature on cost functions for sewer projects. In their appendix, they included an overview table including cost functions of 11 other papers from 1975 till 2010. The following interesting conclusions can be drawn from this table:

- 5 cost functions use interaction terms
- 5 cost functions use non integers powers (example $H^{1.14}$)
- 9 out of 11 calculate price per distance unit
- Only diameter, depth, and length are used as inputs for cost functions (no surrounding factors are considered for any of the projects)
- Only 1 of the cost functions is linear

Drinking water:

Besides sewer systems also natural gas and drinking water systems are quite similar to DH networks. Unfortunately, no literature was found on the cost modeling of natural gas infrastructures, which might be because natural gas infrastructures are less common than drinking water infrastructure on a global scale. However, some interesting papers discussing cost modeling of drinking water systems are discussed below. Marchionni, the same author of the sewer systems paper discussed above, also published two papers on water supply infrastructure cost modeling. The second paper which is an extension of the research presented in the first paper is most relevant for this research. This paper uses multivariate regression based on 130 Portuguese water utility projects to model construction costs of different parts of a water supply infrastructure (Ground storage tank, Elevated tank, Pumping station, Transport pipe, Distribution pipe and Service connection)[46]. For this research the cost functions for the transport and distribution pipes are interesting. It is important to note that out of the 130 considered projects only 91 projects contain transport pipes and 87 projects contain distribution pipes. As for the sewer cost functions mentioned above, different material types have separate cost functions. The six resulting cost functions are shown in Equation 2.17.

Water supply infrastructure Cost Using Regression Techniques[46]:

$$\boxed{\begin{array}{l} \textbf{Transport} \subset \begin{cases} Cm_{Iron} &= 33 + 110 * d_i + 530 * d_i^2 \\ Cm_{PE} &= 30 + 70 * d_i + 960 * d_i^2 \\ Cm_{Steel} &= -886 + 2005 * d_i - 530 * d_i^2 \end{cases} \qquad (2.17a) \\ \textbf{Distribution} \subset \begin{cases} Cm_{Iron} &= -29 + 950 * d_i - 530 * d_i^2 \\ Cm_{PE} &= 22 + 310 * d_i + 300 * d_i^2 \\ Cm_{PVC} &= 29 - 40 * d_i + 470 * d_i^2 \end{cases} \qquad (2.17b) \end{array}}$$

A similar approach as used by Marchionni et al. described above is also applied by Clark et al. already in 2002 [25]. Clark et al. also used multivariate regression to model the construction cost of drinking water networks. The difference, however, is that Clark et al. not only made different cost functions for different material types but they also differentiated different operations in the construction process. They constructed cost functions for: Material cost, Trenching cost, Embedment cost, Valve cost, Dewatering cost, Sheeting and Shoring cost, Horizontal boring cost, Pavement removal, and Replacement cost, and Traffic control cost. For all the above-mentioned costs multiple categories are chosen each with their own corresponding cost function. As can be expected this resulted in a lot of different cost functions that can be used in an additive manner to estimate project cost based on the project properties. All the cost functions have the same general form which is shown in Equation 2.18. For every cost function the corresponding variables a, b, c, d, e and f are estimated using regression. The variable x represent the primary design parameters (most of the time pipe diameter), the variables u is used to differentiate for the categories in a cost function and y is the cost of a certain component. It can be seen that in this regression again a so-called interaction term is used $(x * u)$.

Cost Models for Water Supply Distribution Systems [25]:

$$\boxed{y = a + b * x^c + d * u^e + f * x * u} \qquad (2.18a)$$

Besides using regression, a relatively simple ML approach, to model water supply cost, Shehab et all propose a NN to estimate construction cost of water supply networks [55]. The NN presented in the paper was trained using data from 50 projects that were realized in San Diego, California between 2000 and 2004. The NN was trained using a so-called 80/20 analysis. This method assumes that most of the time 20% of the attributes contribute to 80% of the solution. So for the 50 projects, a set of bid items was identified that contributed to at-least 80% of the total project cost for all projects. The set that was found contained the following 20-bit items that turned out to be the most important (for the 50 projects in San Diego): *"1) new pipes; 2) laterals; 3) abandoned pipes; 4) point repairs; 5) manholes; 6) pedestrian ramps; 7) sidewalks; 8) curbs; 9) gutters; 10) asphalt/concrete pavement; 11) asphalt overlay; 12) resurfacing materials; 13) slurry seal; 14) asphalt patching; 15) excavated soil; 16) imported soil; 17) bedding material; 18) fire hydrants; 19) water services and blow off assemblies and 20) gate valves."* [55]. When looking at the 20 bid items, it can be concluded that a lot of them are related to the surrounding or more specifically to the pavement of the road or road type. The constructed NN indeed generated results that were within 20% of the actual project costs. However, the researchers believe that although the model performed sufficiently, a larger training set would be preferable since it would increase the model's accuracy even further.

In a paper published in 2018, Chee et al. argue that the above-mentioned cost functions, based on historical project data are only valid for high-level planning. These methods are limited because not all pipe diameters and pressure classes for the different types of materials are covered[24]. Also, when the manufacturing process improves different material types not or only scarcely used in the past might become the best solution, predicting cost of "unknown" materials is very challenging using the ML approaches described above by Marchionni [46], Clark [25] and Shehab [55]. Furthermore, in none of the above-described cost functions regional cost adjustment factors can be taken into account. In general, when using historical project data to predict future project cost it is very challenging to make predictions when the proposed project is significantly different from the historical projects used to train the model. [24]. In the paper by Chee et al. an alternative approach is suggested. The Water Pipe Installation COnstruction CoST Estimation (WaterCOSTE) model is presented. The main focus of the model is to calculate the construction cost of very long water transportation pipelines in the united states. In this model, a bottom-up cost building approach based on the process used by contractors to formulate their bidding items is chosen. This approach was chosen based on the assumption that a bid submitted by a contractor is the single best price estimated that can be obtained without realizing the actual project. Chee et al. believe that: "*Cost estimates using regression equations will likely have more variability than those derived from a bottom-up model such as WaterCOSTE*" [24]. The researchers had a hard time proving this statement since no historical project data with enough detailed information was available to validate their model. In the end a comparisons between the WaterCOSTE model and the models presented by Clack[25] and Marchionni [46] was conducted in which WaterCOSTE *"demonstrated valid results"*[24].

2.3. Important conclusions on cost modeling

In this section, the most important conclusions from this chapter are summarized and put into context. The first important conclusion is that the (energy transition) models currently used in professional practice have a large bandwidth. A reasonable part of this uncertainty is probably because these energy transition models currently do not take surrounding parameters into account. Luckily in most of these models, it is possible to enter own cost key parameters. This means that these model could potentially increase their prediction accuracy if more accurate (surrounding dependent) cost key figures would become available.

When looking into the scientific efforts for construction cost predictions of DH networks and similar infrastructures, only a few papers were found and only one of the papers tried to implement surrounding parameters in their cost function. In this paper based on 55 projects, it was concluded that only the presence of asphalt was significant for the project costs.

The two above mentioned conclusions indicate that there is a lot of potential in the field of surrounding based construction cost predictions for DH networks. However, to be able to generate surrounding-based cost key figures a model is required. Based on the literature review in this research it is chosen to use linear regression to develop such a model. This decision is made because:

- For similar cost prediction projects for DH and similar infrastructures linear regression is the most common ML approach to predict cost.
- When data sets are small linear regression seems to work (at least relatively) better [56].
- Linear regression is a lot more interpretable, which is important for this research since a comparison between construction costs of different infrastructures is required.

Besides the decision which ML method to apply in this research also some important conclusions can be drawn on how to use linear regression properly. The three most important conclusions based on the above-presented literature are:

- For linear regression only significant independent variables must be included in the model if too many inputs are used model performance decreases [56]
- Long-term predictions are less accurate than short-term predictions [60], so data should be recent and be updated regularly.
- Interaction terms, in linear regression, seem to be promising for cost predictions of construction cost of infrastructures.

3 Research methodology

In this chapter, the methodology that is applied to answer all the subquestions and finally the main research question, which are presented in section 1.4, is discussed. First, in section 3.1 the (graphical) research design is presented and explained. Second, more information on the model design, and how this model design changed over time, is elaborated on in section 3.2.

3.1. Research design

As already mentioned in section 1.3, the main goal of this research is to develop a statistical model that generates detailed cost parameters (€/m) for DH networks in existing neighborhoods. This model considers surrounding parameters for its cost predictions and uses data from similar infrastructures to increase the amount of potential training data points (realized projects). However, before such a statistical model can be developed also some other questions have to be answered. These subquestions, which are presented in section 1.4, are answered using different research methods. The three main activities which are conducted in this research to answers the research questions are **Literature review**, **Interviews with experts** and **Development of construction cost model**. A graphical overview of the entire research, containing the three research methods (the green outline boxes) mentioned above, is presented in Figure 3.1. As can be seen in the overview the three research approaches are used to solve different kinds of problems (orange filled boxes).

In the literature review, first, it is validated that, currently, there is no model available in the Netherlands that uses surrounding parameters to calculate detailed construction costs for district heating networks. Additionally, it is also checked whether developing a surrounding parameters based construction cost model has the potential to be of added value for policymakers that currently use the available energy transition models. Second, the best-suited modeling approach is identified, considering different types of ML methods. This literature review is described in chapter 2. Afterward based on interviews with experts, infrastructures that are deemed similar enough to DH networks to include in the model are identified. In a new round of expert interviews, surrounding parameters that are likely to influence the construction costs and should therefore be potential model inputs are determined. Furthermore, national GIS databases containing these surrounding parameters have to be found. Finding these national GIS databases is done both by talking to experts and by doing desk research. That is why the GIS data gathering box is not fully in the conducting interviews box. It should be noted that collecting the required GIS databases is done together with colleagues from Deltares also working in the WarmingUP 2C research. Finally, historical project data of all the different considered infrastructures needs to be collected. This data is gathered through a series of interviews with experts from several potential data providing companies. The conclusions from the experts' interviews and the data gathering process is described more elaborately in chapter 4.

After all the 'real' historical project data, with corresponding surrounding parameters from GIS databases, is gathered different sets of self-generated dummy projects are constructed. These dummy projects are generated to validate the modeling approach. Since the dummy project construction costs are based on a fictive self-designed cost function it can be easily validated whether the model can accurately reproduce this known cost function. More information on the different types of dummy data is presented in section 5.4.

Finally, after all the data is gathered and the dummy data is generated the model development can start. The first step in the model development is data preprocessing. In this stage all the gathered project data is converted into a format that can be inputted as training data in the chosen ML model. The data preprocessing of both the project data gathered from companies and the surrounding data gathered from national GIS databases is described in section 5.1. After all the data is converted into a total input table the model performance of the chosen ML model has to be optimized. The optimization of the model is done in two different ways. First, the model is optimized by hand using common so-called feature engineering strategies found in literature. Secondly, an algorithm is designed which combines some of the feature engineering strategies to automatically find the 'best' possible model given the chosen optimization strategy. The main reason why the by hand analysis is conducted, to validate the performance of the algorithm. If the algorithm performs properly it should at least reach the same performance, but it is likely to find a better performing model. When the algorithm finds a model that performs less than the models found in the by-hand optimization approach this indicates that the algorithm is "broken" and should be fixed. The model development process and the design of the algorithm are described in chapter 5.

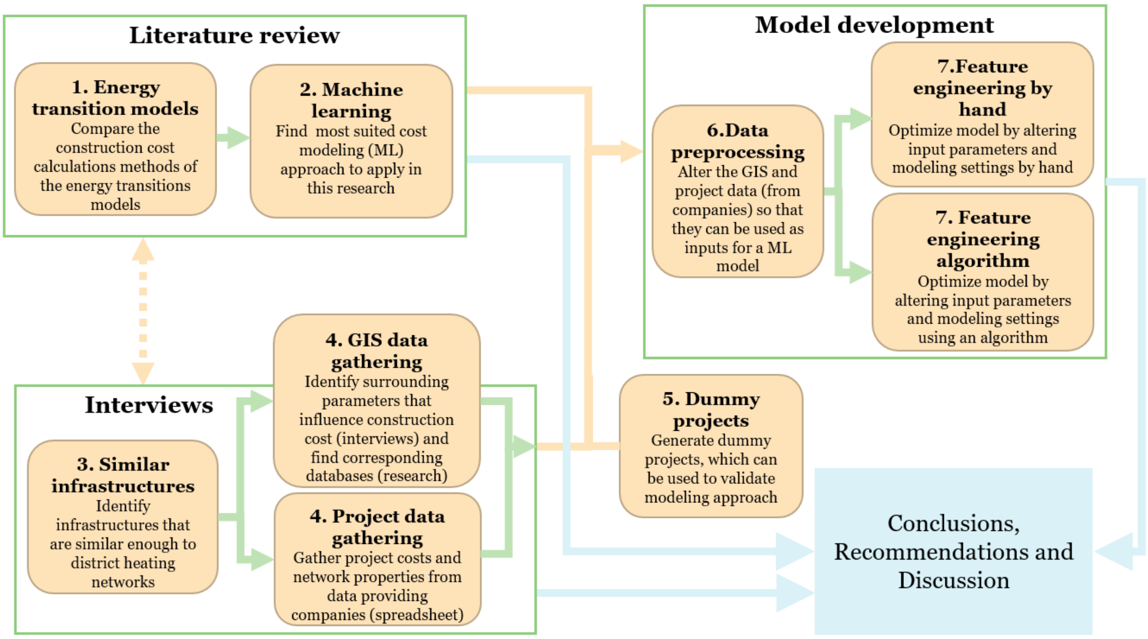

Literature review
1. Energy transition models
Compare the construction cost calculations methods of the energy transitions models
2. Machine learning
Find most suited cost modeling (ML) approach to apply in this research
Model development
6.Data preprocessing
Alter the GIS and project data (from companies) so that they can be used as inputs for a ML model
7.Feature engineering by hand
Optimize model by altering input parameters and modeling settings by hand
7. Feature engineering algorithm
Optimize model by altering input parameters and modeling settings using an algorithm
Interviews
3. Similar infrastructures
Identify infrastructures that are similar enough to district heating networks
4. GIS data gathering
Identify surrounding parameters that influence construction cost (interviews) and find corresponding databases (research)
4. Project data gathering
Gather project costs and network properties from data providing companies (spreadsheet)
5. Dummy projects
Generate dummy projects, which can be used to validate modeling approach
Conclusions, Recommendations and Discussion

Figure 3.1: Research design

There are two final important remarks to be made on the (graphical) research design presented in Figure 3.1. First, as the dotted line between the literature review and the experts' interviews is meant to indicate, the literature review and the interviews were conducted partly in parallel. Even though the literature reviews started earlier, lessons learned in the interviews were used in the literature review and vice versa. The second remark is related to the blue lines with bigger arrowheads. The lines represent information that is used in the conclusion section of the report to answer the subquestions. It can be seen that information from all the research approaches is used to answer the research questions.

3.2. Model design

In this subsection, some more information on the statistical model development is provided. It is important to note that this section focuses on the high over model design. This means that the focus lies on how the data from the different fluid infrastructures can be combined to create a single model that predicts the construction costs of DH. Information on the applied ML approach and the process of training and improving the model performance can be found in chapter 5.

In Figure 3.2 an overview of the original designed model setup is presented. As can be seen in the overview, besides DH, natural gas and drinking water are considered similar enough infrastructures and are therefore included in the research. The reasoning behind the selection process of these two infrastructures is given in section 4.1. The original idea was to generate three separate cost prediction models (circles with number 2.) one for drinking water (blue in overview), one for natural gas (green in overview), and one for DH (orange in overview). As can be seen in the outermost ring of the overview (the circles with number 1.) all three models require the same three inputs. Two of these inputs, the construction cost, and the network properties, are gathered from data providing companies. The third input, surrounding parameters, is gathered from national (GIS) databases using a GIS analysis. After generating three separate models to predict the construction cost of the three similar infrastructures, the final step of the research is to combine these three models into 1 model this potentially can predict the construction cost of DH networks more accurately. In this model, the relations found in the Water and Gas models are also considered. By doing this the total amount of projects on which the final model is based is increased which could increases the prediction accuracy. Of course, it has to be validated that the model performance of the DH model which uses water and gas data outperforms the DH model only using DH data.

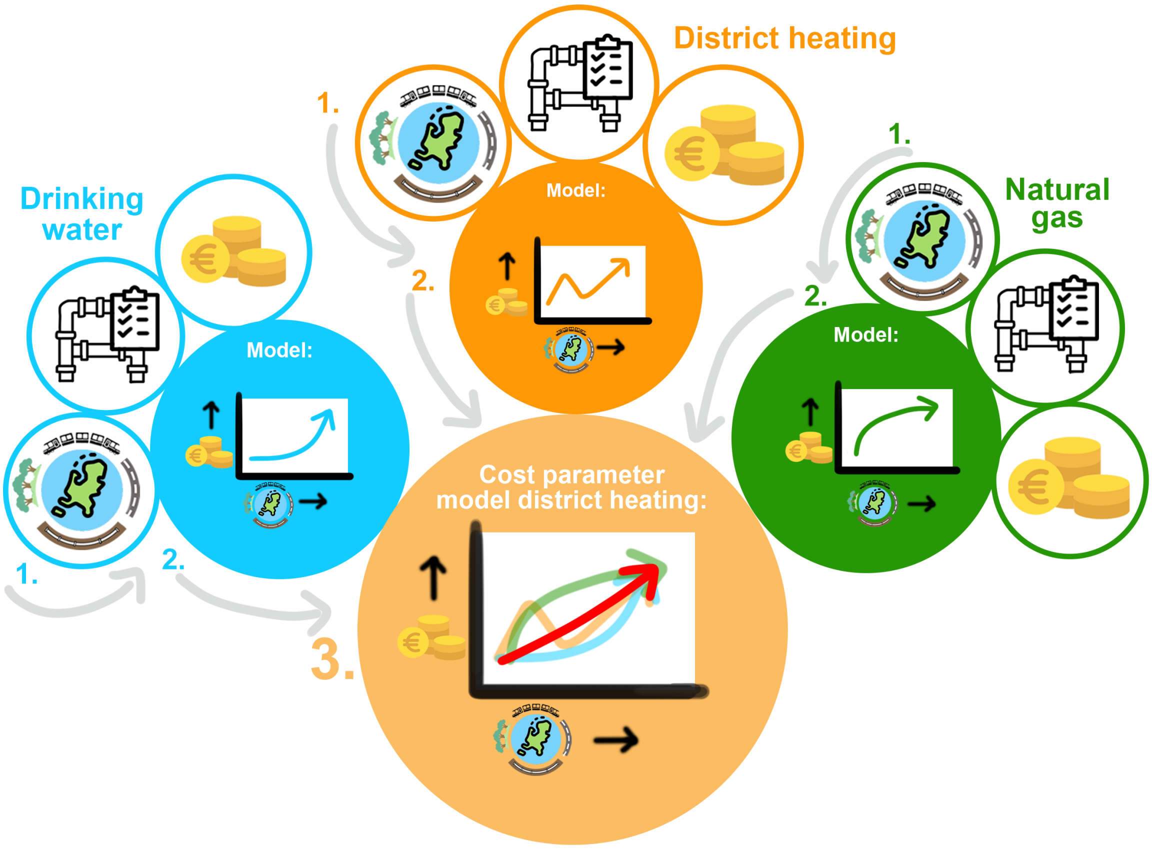


Figure 3.2: Original model overview

However, the gathering of historical project data for DH, which was supposed to happen through the WarmingUp research consortium, turned out to be even more challenging than expected beforehand. This, unfortunately, resulted in a situation where only a significant amount of project data was gathered for Water and Gas replacement projects. This means that the original modeling strategy depicted in Figure 3.2 had to be altered. Even though, this alteration happened in a later stage of the research it has been decided to inform the reader of this change in the methodology section since the rest of the report describes this altered modeling approach. A graphical overview of the altered modeling approach is shown in Figure 3.3. Instead of constructing three models and combining them into 1 'super' model for DH costs, it is decided to develop four models that all have the potential to be of added value for DH predictions. First, the two 'normal' models for water and gas (number 1 and 2 in overview), which were also part of the original model overviews, are developed. Afterward, two combined models are developed that use data from both the water and gas projects to predict respectively water (model number 3) and gas (model number 4) construction costs. This way it can still be validated whether using data from a similar infrastructure is indeed improving the modeling performance. When this is the case for the combined water and gas model it is quite likely that it is also the case for a (future) combined DH model. The final model, model 5.Predicting heat, that is depicted in orange, is not developed in this thesis. However, the WarmingUP 2C research team is still planning on developing the final heat predicting model in the near future. For that reason, this research is focused on providing useful inputs and conclusions for that future model. Since the main focus of the thesis still lays on surrounding based construction costs predictions of DH networks in the built environment.

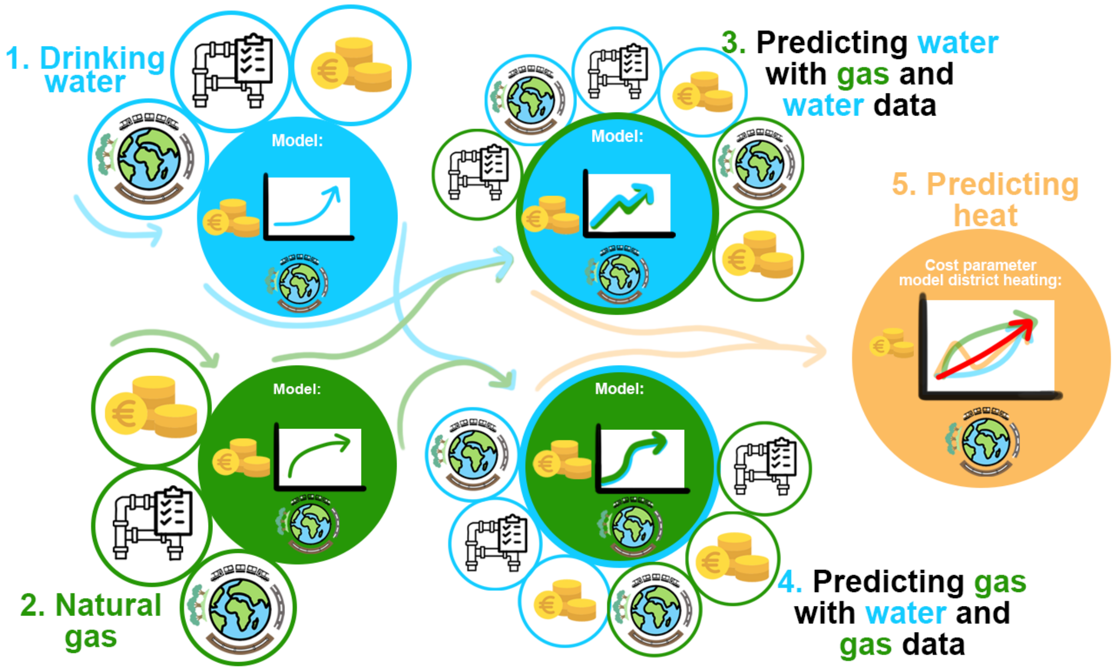

1. Drinking water
Model:
3. Predicting water with gas and water data
Model:
5. Predicting heat
Cost parameter model district heating:
Model:
Model:
2. Natural gas
4. Predicting gas with water and gas data

Figure 3.3: Updated model overview

Unfortunately, the resulting Water, Gas, and combination models are no longer directly applicable to DH networks. Nevertheless, the results of the water, gas, and the combined models can still be of added value for the DH sector in the three following ways:

- The study can provide a good proof of concept for construction cost modeling considering surrounding parameters. If the results of this study are promising, and when in the future a sufficient database with historical DH projects is realized, the method applied in this thesis can be used again. This time resulting in a model that can be used specifically to predict the construction cost of DH. It is noteworthy to mention that the WarmingUP 2c project team is planning on realizing a database with historical DH projects and using this to model DH cost in the near future.

- Since, drinking water, natural gas, and DH networks are relatively similar, it is quite likely that the significant relations found in the Water and Gas model are also applicable up to some extent for DH construction cost estimations. The potential use of the developed models for DH predictions and some critical remarks on the limitations of the models are presented in section 6.6 and chapter 7 respectively.

- The data gathered in this research can potentially be used as extra training data for a DH heating ML model.

II

Part 2 Conducted research

4 Interviews and Data Gathering

There is very limited literature available on surrounding based cost modeling of fluid infrastructures and no literature is found on the similarities and differences between different fluid infrastructures in the Netherlands. Therefore, a big part of this research was realized by talking to domain experts in the fields of cost modeling, GIS, district heating, drinking water, and natural gas. In this chapter, the interviews that were conducted as a contribution to this research are discussed.

The chapter is separated into two parts. The first part focuses on similar infrastructures. In section 4.1, first, all the infrastructures that are deemed similar enough to DH to include in this research are identified. Afterward, the similarities and differences between these chosen infrastructures and DH networks are described. These descriptions are predominantly based on interviews conducted with domain experts. The second part describes the historical project data gathering process in section 4.2. This process mostly consists of talking to companies to identify which data is available and under which conditions they are willing to share this data. An exception is the gathering of surrounding data which is mostly collected from national open-source GIS databases instead of companies. However, the considered surrounding parameters that are gathered are based on experts' opinions and this process is therefore also dependent on the conducted interviews.

4.1. Similar infrastructures

The goal of this research is to find relations between important surrounding factors and the construction cost of DH networks. However, as already mentioned in chapter 3, the analysis doesn't use historical data from DH networks, but instead historical data from similar infrastructure projects is used. The question that remains is which infrastructures are similar enough to DH networks to conduct such an analysis.

The question of which infrastructures are similar enough to DH networks to be able to compare cost dependency on surrounding factors is a tradeoff between the number of projects which can potentially be used and the similarity of the construction process of the different infrastructures. The first attribute that was considered for this tradeoff is the construction speed (meters/hour). If the construction speed of an infrastructure is low the project will take longer and the project cost will be higher. When considering construction speeds, the conclusion can be drawn that only pipe-based infrastructures are promising. Since cable infrastructures, like electricity and internet cables, have a construction speed which can be more than 100 times higher, compared to the pipe infrastructures. The pipe-based infrastructures which are left for consideration, i.e. drinking water, natural gas, and sewer systems, are analyzed more elaborately. The similarities and differences of these infrastructures compared to DH are identified through interviews with domain experts. In one of the first interviews, sewer systems were discarded. This decision was made because, in contrast to water, gas, and heat infrastructures, sewer systems do not use pumps (in the built environment) but gravity as a driving force to move the fluid inside. The consequence of this is that for sewer systems it is significantly harder to move around obstacles, like other pipes and roots in the vertical direction, during the construction. Because of this reason, it was decided that sewer systems were not suitable for this analysis.

So the infrastructure that were deemed similar enough to DH networks for this analysis are drinking water and natural gas networks. Even though these networks are similar to DH networks they do of course differ in some aspects. To be able to compare the cost of these different infrastructures it is of high importance to be aware of these differences. Therefore, the differences of the three infrastructures are discussed in the next tree subsection. Thereafter, the similarities are mentioned. The differences discussed below are all conclusions from 34 interviews that were conducted with domain experts on drinking water, natural gas, district heating networks, and spatial planning. An overview of all the interviews that where conducted, including the 34 relevant for this subsection, can be found in Table B.1 in Appendix B

4.1.1. Differences in pipes:

First, let us consider the differences in the pipes of the different networks. The first clear difference is the amount of pipe that is required. Since DH is a closed-loop system it contains a supply and return pipe meaning that twice the amount of pipe is required when connecting the same amount of houses using the same trace. The only exception for this is when a so-called multi-pipe is used where the supply and return pipe are incorporated in the same pipe see Figure 4.1. Moreover, current DH networks most of the time consist of steel pipes whereas most (new) gas and water networks consist of PE and PVC pipes. Additionally different pipe materials of the different infrastructures, sometimes need to be connected in different ways. Currently, most district heating networks consist of steel pipes, meaning that the different pipes need to be welded on site. Since additional space is required for the welding process DH network with steal pipe require trenches that are deeper and wider compared to PVC or PE trenches. Also, welding generally is more time consuming that the clicking systems commonly used for the PVC and PE pipes. Keep in mind that most new, low temperature, DH networks also use PVC or PE pipe instead of steal. Furthermore, DH pipes are isolated meaning that their outside diameter will increase and maintenance requirements change (keep in mind that maintenance cost are outside of the scope of this thesis). Finally, when specifically comparing natural gas and DH, DH pipes have a bigger diameter when they need to provide the same amount of thermal energy. This is due to the lower energy density of hot water compared to natural gas.

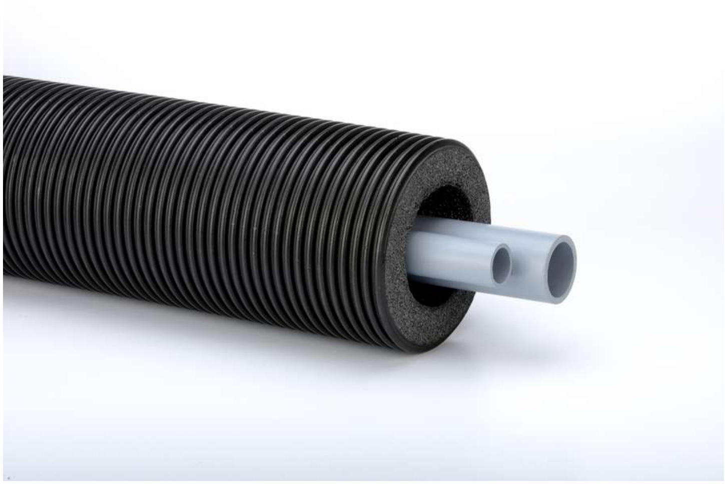


Figure 4.1: Pre-insulated Multi pipe Flexalen 1000+ from Thermaflex [4]

4.1.2. Differences in construction process:

First of all, it is important to note that there is a difference in the type of projects that are considered in this research for heat networks compared to gas and water networks. For heat networks, projects where new networks are realized in the built environment, are considered. Whereas, for gas and water, projects, where old pipes are replaced by new pipes (again in the built environment), are studied. This difference in project type results in the first two differences between the DH and water and gas construction process. First of all, for water and gas projects not only new pipes have to be installed but also, for most projects, old pipes have to be removed (some water utility companies leave old infrastructure temporarily in place as back up). This can lead to different project cost for multiple reasons. Some examples are: an extra trench has to be dug when old and new pipes are not in the same place, the trench for construction has to be deeper than necessary because the old pipes sagged over time and the old pipes have to be transported to a waste facility. A second difference when replacing water and gas compared to constructing a new district heating network is that during a water and gas reconstruction project connected households should still be able to use gas and water. Whereas in most district heating project houses are still connected to the gas grid during the construction phase. The fact that houses need their gas and water during the project complicates the project and could lead to higher costs. To solve this most of the time projects are subdivided into small batches that can be realized in a day. Meaning that connected household do not have water and or gas during the day but will have access again in the evening and at night.

Furthermore, certain additional rules regarding safety are in place for gas and drinking water construction projects. For drinking water for example after the construction is finished everything has to be flushed with chlorine and a validation test is required to make sure that the hygiene standards are met. This cleaning and validation process could take approximately two extra days. A final important aspect is the possibility to do combi-work when maintaining or installing subsurface infrastructures. It is quite common in these types of projects to work together and use the same trench for multiple infrastructures. Note that in Amsterdam for example it is sometimes even mandatory to work together with different infrastructures. The problem here, when trying to predict costs, is that there are different (pre-defined) rules for dividing costs between the different infrastructures. This is because different infrastructures have different requirements for the trench, for example: depth, width, the time required in the trench but also the urgency to dig the trench now rather than next year. Since these factors can influence the cost significantly the costs are divided accordingly. Meaning that the same trench can have a different cost for a certain stakeholder in different projects depending on which other infrastructures were participating in a particular project. This is important to keep in mind when comparing historical costs from different projects from both the same and different infrastructures.

4.1.3. Differences in network design and scale:

Water and gas networks are designed differently compared to heating networks. Water and gas are designed less "efficient", meaning that they use more meters of pipe per connection in similar neighborhoods. They do this so the system is more resilient, if a certain pipe fails somewhere other pipes will take over their demand. It is important to note that this is not very important when calculating the construction cost per meter. Also, water networks in contrast to gas and heat are not dimensioned based on the demand of connected houses but the required pressure for fire fighting water is the design limitation.

Furthermore, the location of the pipes can also be different for the different infrastructures. Drinking water prefers to be on the shadow side of the road whereas heating would prefer the sunny side. Also drinking water and DH should not be to close to each other to minimize heat transfer between the pipes. Especially where water and heat enter the house this can be challenging. Additionally, some municipalities have rules for the location of specific infrastructures below roads and sidewalks. Generally speaking new DH networks have to be placed below the street and drinking water and natural gas are placed below the sidewalks. Since sidewalks are closer to the houses and generally are not made of asphalt placing your infrastructure below the sidewalk is cheaper and therefore most desirable. However, the subsurface below sidewalks in (big) cities is getting full. Since, as already mentioned above, district heating requires more space than water and gas most DH networks are "banned" to the subsurface below the street. An interview about specifically this topic, location in the subsurface, was conducted with an urban planner of the municipality of Amsterdam. In this interview, it was validated that indeed in most of their future street profile designs DH networks are placed below the street. Since contractors require permission from a municipality before they can construct a new network it is quite likely that indeed most future heat networks will be placed below the street.

Finally, the scale of the networks is different, where water and gas networks cover almost the entire country and are therefore are generally quite big. Most district heating networks, in The Netherlands, are smaller. Meaning that the distance between the source and the households and is generally smaller for heat networks than for water and gas. A transport pipe for an average heat network would probably be considered part of the distribution network according to the gas and water standards. This means that even though the same terms (transport, distribution, and connection pipes) are used in the different infrastructures, these terms will not necessarily have the same characteristics. Because the scoping of the research is not based on the network type and since the considered water and gas replacement projects are relatively small. This difference in size of the total network does not cause significant differences between these networks for this research.

4.1.4. Similarities gas, water and heat:

Despite the differences, between the three infrastructures that are discussed above, the general consensus after conducting the interviews with domain experts is that it is indeed promising to compare the cost dependency on the surroundings for the three infrastructures. The experts that were interviewed feel that there is a considerable overlap in the project's cost, and more importantly, they feel like the project cost is dependent on the surroundings in a very similar matter. This is because the construction methods (open trench, drilling, sinkers, etc) and used materials (PE, PVC and steal) are similar. Also, the same contractors with similar contracts are hired to construct the networks. In short, most experts felt like the total project costs (also per meter) are probably different (due to material cost, different deals for combi-work, different safety measures, etc). But most of them feel like it is promising to research the relation between the cost and limiting surrounding parameters for the different infrastructures since they are likely to be similar.

4.2. Data gathering historical projects

In this subsection, the data gathering process is discussed. For this research data is gathered for drinking water and natural gas replacement projects in the built environment in The Netherlands. As already mentioned in chapter 3 originally the goal was to also collect historical project data for district heating projects through consortium partners of the Warming UP research consortium. However, due to some difficulties with commutation in the consortium itself and the higher-level managers of the partner companies, the process of data gathering of heating projects is seriously delayed. This, unfortunately, resulted in the situation that no heating data was available in time, to be used for the analysis conducted in this thesis. Since describing the process of collecting data which in the end did not succeed is deemed irrelevant for this thesis the rest of this subsection focuses on the data gathering from Water and Gas data only. It is, however, important to note that the data gathering method, using a data-gathering spreadsheet and GIS databases, is the same for water and gas and was supposed to be very similar for DH data.

As can be seen in the research overview depicted in Figure 3.3 in chapter 3 three different types of inputs are required for all the ML models. The three different inputs that are required are project costs, network properties, and surrounding parameters. As can also be seen in Figure 4.2 the first two, project costs and networks properties, are gathered using the data-gathering spreadsheet from water and gas utility companies, and the surrounding parameters are gathered from national open-source GIS databases. In the rest of this subsection, the gathering of these three different types of data is discussed. First, in subsection 4.2.1 the data providing companies are presented. Second, in subsection 4.2.2 the method for data gathering using the data-gathering spreadsheet is described and afterward in subsection 4.2.3 the GIS based approach for gathering surrounding parameters is discussed.

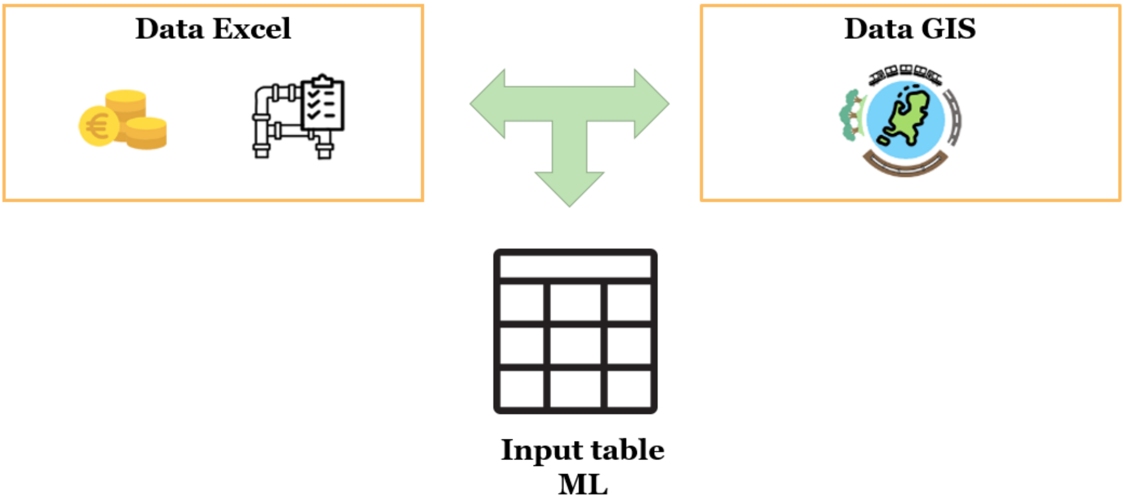


Figure 4.2: Data gathering overview

4.2.1. The data providing companies

To analyze data from historical projects this data needs to be gathered first. However, before historical project data can be gathered it is important to identify who can provide the required data and what specific data is required in the first place. In other words, it is important to first talk to potential data suppliers to identify what kind of data they have saved in their systems and under which conditions they are willing to share that data. The two logical parties to talk to when you are interested in Water and Gas data are the Water and Gas utility companies and the contractors who realized the projects. After a couple of interviews, two things became clear.

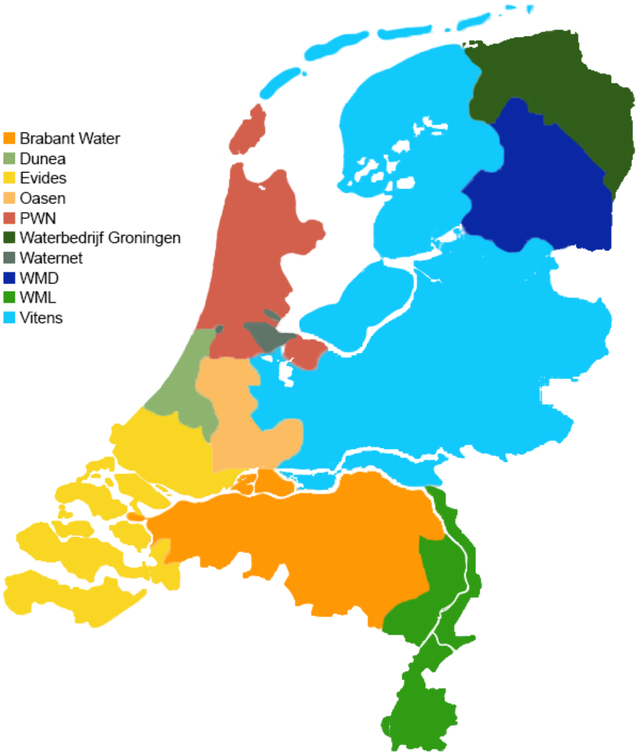


Figure 4.3: Overview water utility companies in The Netherlands [14]

First, the most detailed cost data was available in the contractors' cost databases. Since the water and Gas utility companies are most of the time not present during the actual construction process. This originally resulted in a preference for using contractors as data suppliers. However, the second realization, namely cost data is highly confidential for contractors, changed this perspective. It became clear that contractors were not willing to provide cost data on such a high detail level and were quite skeptical to share any data at all. After two interviews with contractors and some advice from consortium partners, the conclusion was drawn that contractors do not gain enough added value from this research, so they are probably not willing to take a big risk by sharing cost data. Especially since all contractors are commercial companies, and are therefore less likely to participate for "the greater good", the decision was made to change the focus on the semi government, water and gas utility companies.

Project data water utility companies:

In the Netherlands 10 different water utility companies exist. All the water utilities have their own service area see Figure 4.3. This means that water utilities are no rivals to one another since their customers can not choose a different provider. A large advantage of this, compared to the contractors, is that cost data for water utilities is a lot less sensitive. Also, water utilities are already used to working together and providing data for this purpose. A good example of this is the "kostenstandaard drinkwater", this cost calculation tool developed by Royal HaskoningDHV is used by 90% of the dutch water utilities for their investment plans. In the Netherlands therefore this is the leading model with respect to cost predictions of drinking water infrastructures, this is also validated through interviews that were conducted with all water utilities. The tool has cost functions for water production, water purifying, water storage and water transportation [53]. All the water utilities with a license pay 0.006% of their yearly revenue and keep providing new data, or at least they are supposed to. With this money and the new data, Royal Haskoning keeps the tool up to date. An interview with the product manager of the **"kostenstandaard drinkwater"** let to a couple of interesting conclusions. The most interesting conclusions are:

- The tool does consider surrounding effects on the construction cost of transport pipes but does so by category (rural, urban, in-between).
- Gathering historical project data is challenging, it is very important that the amount of time that is

required for a company to provide the data is as low as possible. When the required time is too long the amount of data that is provided reduces significantly.

- The tool uses the least-squares method for predicting cost based on old projects.

- The tool originally only focused on water purification costs but now costs for transport and distribution pipes are also available. Especially in the distribution model, there is some room for improvement since there is not that much data available yet.

- The tool can also be used for cost calculations of similar infrastructures (hydrogen, district heating, CO_2), which indicates that indeed similar infrastructures have similar cost structures.

- The goal of the tool is to realize a bandwidth of 30%, however, in the construction cost modules of the tool this is not yet realized.

In the first interview with the first water utility company, a list was provided with contact details from asset managers from all the ten water utilities in The Netherlands. After a lot of emails, in the end, all water utilities agreed to plan a meeting to talk about this research. In the interviews, four main discussion topics were addressed. For an overview of all the interviews that were conducted see Table B.1 in Appendix B. The four topics that were discussed are:

- Similarities and differences between water, gas and DH networks

- Surrounding factors that could potentially influence the construction costs

- Identifying the types and the amount of data that was available on historical replacement projects.

- utility companies were asked whether they were willing to share historical cost data and if so how they would prefer to do it

The first point of discussion was meant to identify the possibilities and challenges of using water and gas data for DH predictions. The important conclusions from this part of the interviews are already discussed in section 4.1. The important conclusions from the second discussion point can be found in subsection 4.2.3 where the gathering of these important surrounding parameters from national GIS databases is discussed.

From the third discussion point, it became clear that the detail level of data that is available is less than was expected. Also, it turned out that the file systems of the different water utilities were not as similar as expected beforehand. Meaning that the amount of data that was available at every single utility was relatively limited. Finally, some of the water utilities pointed out that even though the cost data might be available on a reasonable detail level, that would not necessarily mean that they could also share this data. This is because even though the data might not be of competitive value for the water utilities it is still sensitive information for the partner contractors that realized the projects. A water utility said that they could only share certain cost data with the permission of their contractors. Again realizing that contractors are not very likely to give that permission the conclusion was drawn that the cost data should be gathered on a sufficiently high aggregation level so that water utilities can choose for themselves whether they would share the data.

The fourth discussion point was for most water utilities discussed in a second interview round when the first draft of a data-gathering spreadsheet was constructed based on the already conducted interviews. More information on the data gathering spreadsheet can be found in subsection 4.2.2. In the end, four out of the ten water utilities were willing to cooperate with the research by sharing project data. The six other utilities either did not have sufficient data in their systems and/or did not have time available to gather project data as input for this research. Two out of the four utilities did not have enough time to fill the data gathering spreadsheets themselves so those utilities decided to send all the required project information directly from their file systems. For those two utilities, the required information had thus be found in the sent documentation and be filled into the data gathering spreadsheet by hand. For all utilities, the first two projects were done together during a (MS teams) meeting to make sure that all the information was interpreted correctly. An overview of the water replacement projects that are gathered can be seen in Table 4.1.

Table 4.1: Data gathered water utility companies

Name	spreadsheet	Project data	Number of projects
Utility 1	x		16
Utility 2	x		13
Utility 3		x	32
Utility 4		x	9

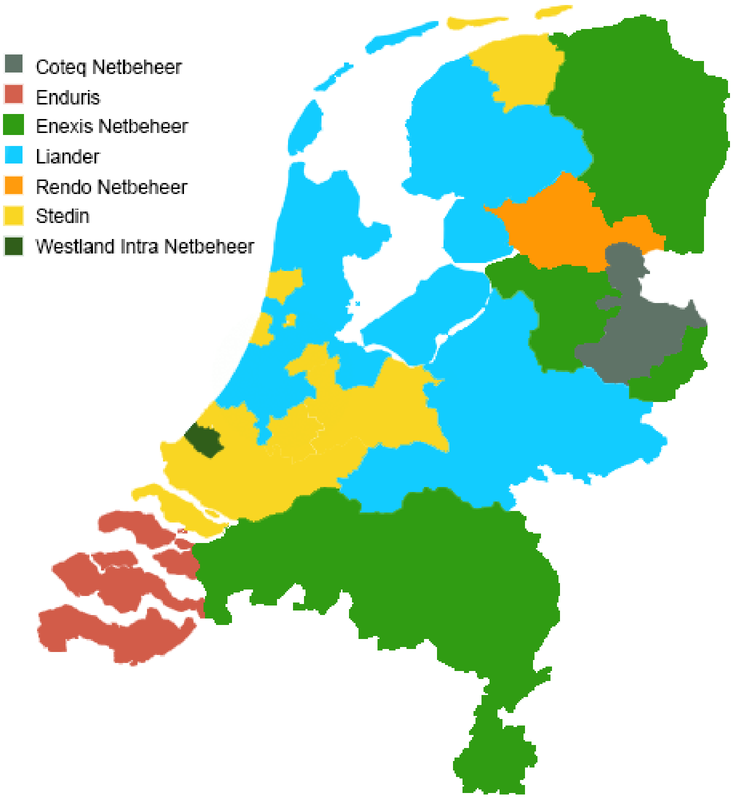


Figure 4.4: Overview network operators in The Netherlands [13]

Project data gas utility companies:

In the Netherlands, 7 network operators together manage the natural gas distribution network of the entire country. However, as can be seen in Figure 4.4 three of these network operators (Liander, Stedin, and Enexis) together account for more than three-quarters of the country. That is why, also taking into account the amount of effort and patience it took to get appointments with the ten water utilities when the contact details were already known, it was decided to only reach out to these three big network operators for cost data. Luckily all the three network operators, in the end, agreed to talk and see if they could help with the research. In the interviews with the network operators the same four discussion points as described for the water gathering process in Figure 4.2.1, where addressed. The similarities and differences and the important surrounding parameters again are already described in section 4.1 and subsection 4.2.3 respectively. The same issues with the amount of data that was available and the detail level of the cost data that were faced with the water utilities were also faced when talking to the Network operators. Fortunately, in the end, all network operators agreed to help with the research and provide historical project data. As can be seen in Table 4.2 this time again two utilities chose to share project data instead of filling the data gathering spreadsheets themselves. Again as for the water spreadsheets, the first two spreadsheets for all utilities were filled together to make sure that everything was interpreted correctly.

Table 4.2: Data gathering network operators

Name	Spreadsheet	Project data	Number of projects
Network operator 1		x	18
Network operator 2	x		18
Network operator 3		x	16

4.2.2. The data gathering spreadsheet

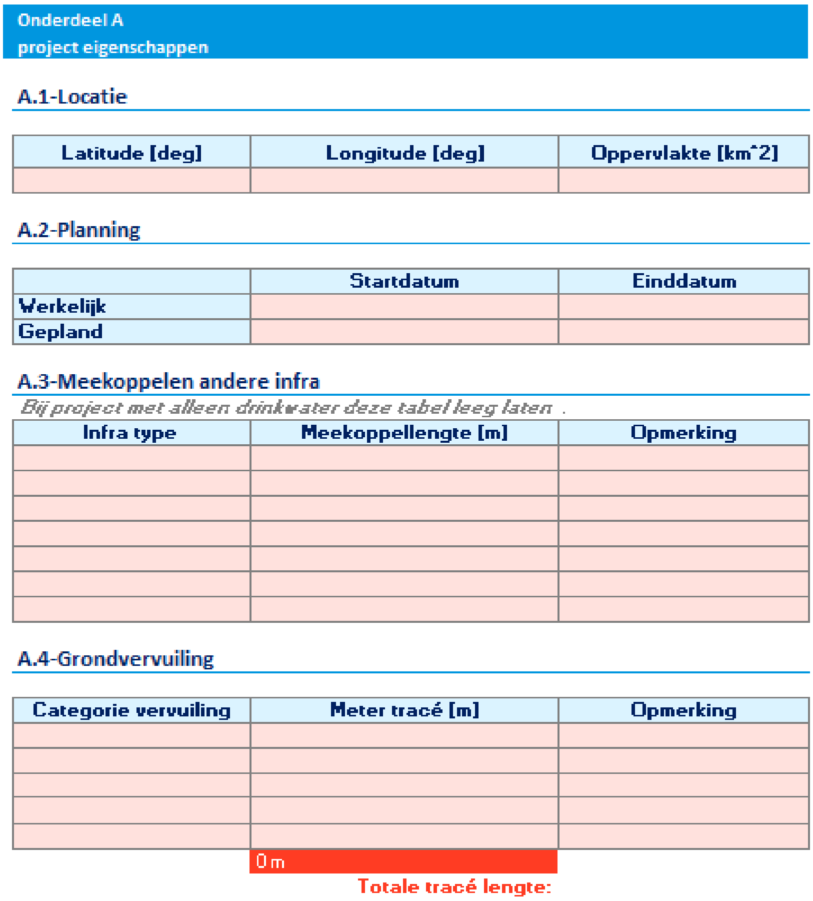
As mentioned before a spreadsheet was used to gather data from utility companies. All the spreadsheets that are gathered are loaded into Python and are merged with other spreadsheets from the same infrastructure. This way two total input tables are generated. One table with all the project data from the water utilities and one table with all the project data from the gas utilities. The question that remains is, what kind of data is actually gathered from the company and what does the data gathering spreadsheet and the total tables look like. Those questions will be answered in this subsection.

During the design of the spreadsheet two tradeoff had to be made. The first was, what should the spreadsheet look like. Things like, how many different sheets should be used, how much context explanation is required, how much explanation on the required inputs is needed and how can these explanations be given in the best way. When more sheets and or explanations are used the context and the requirements for this project become more clear and therefore the gathered data is likely to be of higher quality. However, when too many details are given the required time to fill the data gathering spreadsheet increases significantly which is undesirable since companies are less likely to cooperate.

The second important aspect is the trade-off for the amount and the types of data that are asked for in the spreadsheet. When asking for too much (detailed) information, data providing companies are less likely to cooperate since it will cost them to much time and/or the data might be confidential. However, when too little data is gathered the analysis might not work properly since certain important parameters can not be used as inputs for the model.

For the two above described trade-offs, a decision was made based on interviews with the water and gas utilities and a couple of iterations with feedback rounds from partners (heating companies and contractors) in the Warming UP research consortium. Finally, the data-gathering spreadsheet was constructed within WarmingUP project 2C based on the lessons learned in this research from the water and gas utilities and the lessons learned in the consortium about available data and requirements for DH projects. Generally speaking, for the project and network properties (part A and B spreadsheet), everything that is available, for at-least some companies (heat, water, or gas), and has to potential to influence the cost is included in the spreadsheet. When a certain company does not possess the required data parts of the spreadsheet can be left empty. For the project costs (part C spreadsheet) costs were requested on a relatively high detail level to make sure that the cost data is no longer considered sensitive information by the underlying contractors.

This first draft was designed in such a way that with only minor textual changes the same spreadsheet could be used for Water, Gas, and Heat data collection. The reasoning behind this was that when the same spreadsheet was used for the three different infrastructures, the same data with the same potential bias from miss interpretations in the spreadsheet would be collected for all infrastructures. Also loading the spreadsheet into Python and performing alterations is more convenient if the spreadsheet is very similar for all different projects. The resulting spreadsheet is subdivided into three different parts which are described in the rest of this subsection.



Onderdeel A
project eigenschappen

A.1-Locatie

Latitude [deg]	Longitude [deg]	Oppervlakte [km^2]

A.2-Planning

	Startdatum	Einddatum
Werkelijk		
Gepland		

A.3-Meekoppelen andere infra

Bij project met alleen drinkwater deze tabel leeg laten .

Infra type	Meekoppellengte [m]	Opmerking

A.4-Grondvervuiling

Categorie vervuiling	Meter tracé [m]	Opmerking

0 m
Totale tracé lengte:

Figure 4.5: Screenshot data gathering spreadsheet part A: project properties

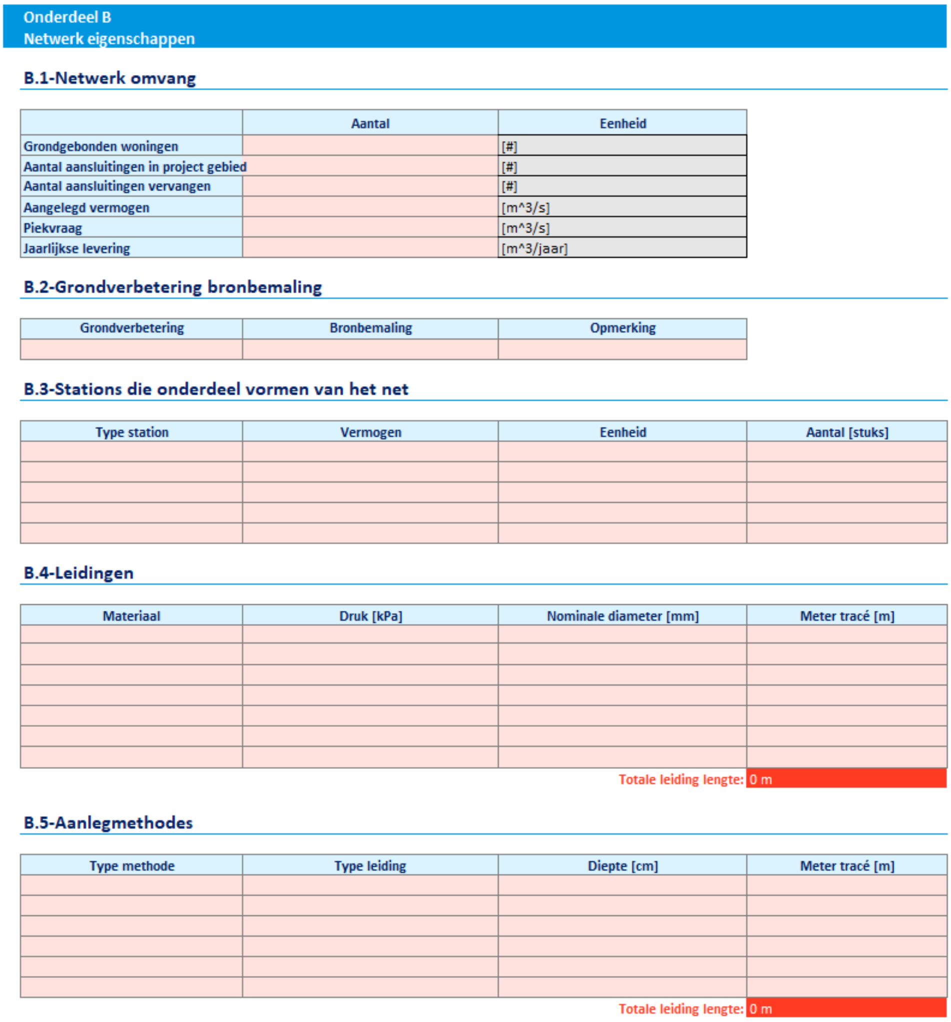
In the first part, which can be seen in Figure 4.5 companies have to provide project properties. In the first three red boxes the **latitude, longitude, and project area** can be filled in case no trace information is available. Luckily for all gathered projects, more information about the trace was available in either GIS or PDF format so these boxes remained empty.

Below, in the "A.2-planning" part, information on the duration of the projects is requested. However, most companies gave warnings that the **start and finish dates** of their project in the file systems are not very reliable and that it is quite common that dates are empty or guessed in a later stage.

In the third table, "A-3-Meekoppelen andere infra", a company needed to indicate whether a project was realized together with another infrastructure, so-called **combi-working**. Combi-working is quite common in these infrastructure projects since the cost for trench digging can be subdivided over different companies. Since this could reduce the project cost significantly it is important to know in the analysis whether other infrastructures were included in the project.

In the final table of part A, **ground pollution levels** are requested. This surrounding parameter is asked in the spreadsheet because no national database containing ground pollution levels was found. When provided in the spreadsheet it is still possible to check whether ground pollution level does affect the construction costs.

In the interviews with gas and water utilities, most utilities indicated that they did not store any information about the surrounding parameters except the ground pollution level. For this reason no other data on the surrounding is asked for in the spreadsheet. The pollution level is available in most databases because utilities have to know this beforehand to implement required safety measures when the pollution levels are severe. These safety measures are also the reason why ground pollution affects the construction costs and is thus potentially a useful input for this analysis.



Onderdeel B
Netwerk eigenschappen

B.1-Netwerk omvang

	Aantal	Eenheid
Grondgebonden woningen		[#]
Aantal aansluitingen in project gebied		[#]
Aantal aansluitingen vervangen		[#]
Aangelegd vermogen		[m^3/s]
Piekvraag		[m^3/s]
Jaarlijkse levering		[m^3/jaar]

B.2-Grondverbetering bronbemaling

Grondverbetering	Bronbemaling	Opmerking

B.3-Stations die onderdeel vormen van het net

Type station	Vermogen	Eenheid	Aantal [stuks]

B.4-Leidingen

Materiaal	Druk [kPa]	Nominale diameter [mm]	Meter tracé [m]
		Totale leiding lengte:	0 m

B.5-Aanlegmethodes

Type methode	Type leiding	Diepte [cm]	Meter tracé [m]
		Totale leiding lengte:	0 m

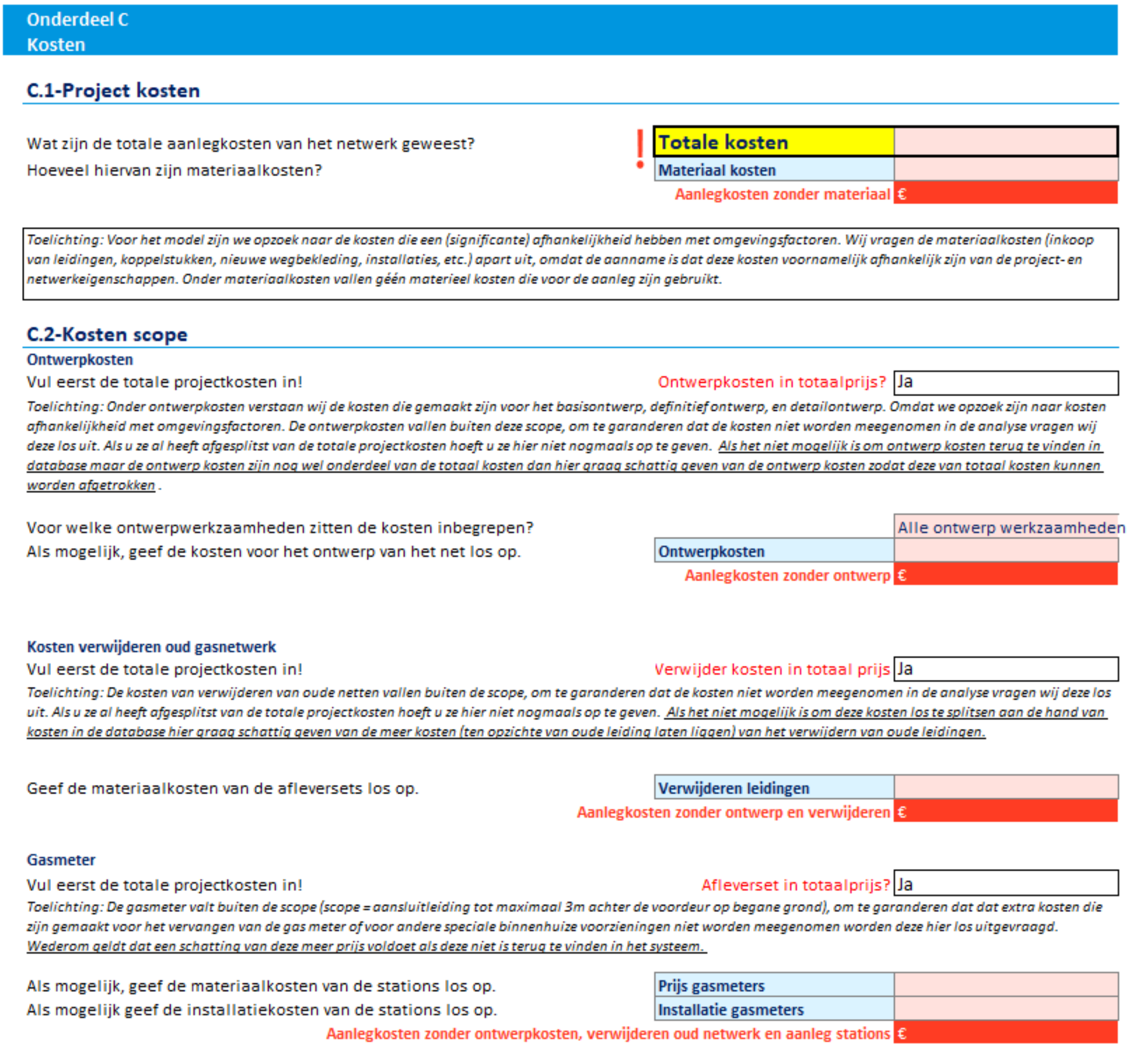
Figure 4.6: Screenshot data gathering spreadsheet part B: network properties

In the second part, shown in Figure 4.6, companies can insert network properties. In the first table information about the size of the network can be entered. However, after talking to the water and gas utilities and showing them the spreadsheet, it turned out that yearly delivery and the rated power of the network were not available and also not deemed relevant. An input that is relevant in this table is the amount of **replaced connections**. In the water and gas replacement projects per default, house connection pipes are not replaced. When in a project some or all of the connections are replaced this can influence the construction cost per meter since most construction pipes are in gardens instead of public space.

The second table, "B.2 grondverbetering bronbemaling" contains two yes or no questions. namely whether or not the **ground is replaced** and whether or not **dewatering** was applied. These questions are included in the spreadsheet to check whether it is possible to base these activities on national GIS layers. For dewatering, you would expect that project where the average groundwater level is high are more likely to have applied dewatering. For ground replacement, you would expect something similar but then for projects where clay is present in the subsurface. Pipes in clay will sag more than pipes in sand, so when a lot of clay is found this is replaced by sand to prevent this sagging.

In the third table information about the (pumping) stations requested. This table, however, is also irrelevant for water and gas projects since the stations are replaced separately from the distribution pipes and the cost for this replacement is therefore not included in this analysis. This actually is a good thing since the water and gas stations are quite different when compared to the HTS stations found in DH networks.

In the final two tables the **pipe diameter, pipe material, construction method**, and the **amount of meters** are required. As can be seen, multiple different material types and or construction methods can be filled in the spreadsheet. Since bigger projects can exist out of multiple material types or diameters and can be constructed in different ways.



Onderdeel C
Kosten

C.1-Project kosten

Wat zijn de totale aanlegkosten van het netwerk geweest?
Hoeveel hiervan zijn materiaalkosten?

Totale kosten	
Materiaal kosten	
Aanlegkosten zonder materiaal	€

Toelichting: Voor het model zijn we opzoek naar de kosten die een (significante) afhankelijkheid hebben met omgevingsfactoren. Wij vragen de materiaalkosten (inkoop van leidingen, koppelstukken, nieuwe wegbekleding, installaties, etc.) apart uit, omdat de aanname is dat deze kosten voornamelijk afhankelijk zijn van de project- en netwerkeigenschappen. Onder materiaalkosten vallen géén materieel kosten die voor de aanleg zijn gebruikt.

C.2-Kosten scope

Ontwerpkosten

Vul eerst de totale projectkosten in!

Ontwerpkosten in totaalprijs? Ja

Toelichting: Onder ontwerpkosten verstaan wij de kosten die gemaakt zijn voor het basisontwerp, definitief ontwerp, en detailontwerp. Omdat we opzoek zijn naar kosten afhankelijkheid met omgevingsfactoren. De ontwerpkosten vallen buiten deze scope, om te garanderen dat de kosten niet worden meegenomen in de analyse vragen wij deze los uit. Als u ze al heeft afgesplitst van de totale projectkosten hoeft u ze hier niet nogmaals op te geven. Als het niet mogelijk is om ontwerp kosten terug te vinden in database maar de ontwerp kosten zijn nog wel onderdeel van de totaal kosten dan hier graag schattig geven van de ontwerp kosten zodat deze van totaal kosten kunnen worden afgetrokken .

Voor welke ontwerpwerkzaamheden zitten de kosten inbegrepen? Alle ontwerp werkzaamheden
Als mogelijk, geef de kosten voor het ontwerp van het net los op.

Ontwerpkosten	
Aanlegkosten zonder ontwerp	€

Kosten verwijderen oud gasnetwerk

Vul eerst de totale projectkosten in!

Verwijder kosten in totaal prijs Ja

Toelichting: De kosten van verwijderen van oude netten vallen buiten de scope, om te garanderen dat de kosten niet worden meegenomen in de analyse vragen wij deze los uit. Als u ze al heeft afgesplitst van de totale projectkosten hoeft u ze hier niet nogmaals op te geven. Als het niet mogelijk is om deze kosten los te splitsen aan de hand van kosten in de database hier graag schattig geven van de meer kosten (ten opzichte van oude leiding laten liggen) van het verwijdern van oude leidingen.

Geef de materiaalkosten van de afleversets los op.

Verwijderen leidingen	
Aanlegkosten zonder ontwerp en verwijderen	€

Gasmeter

Vul eerst de totale projectkosten in!

Afleverset in totaalprijs? Ja

Toelichting: De gasmeter valt buiten de scope (scope = aansluitleiding tot maximaal 3m achter de voordeur op begane grond), om te garanderen dat dat extra kosten die zijn gemaakt voor het vervangen van de gas meter of voor andere speciale binnenhuize voorzieningen niet worden meegenomen worden deze hier los uitgevraagd. Wederom geldt dat een schatting van deze meer prijs voldoet als deze niet is terug te vinden in het systeem.

Als mogelijk, geef de materiaalkosten van de stations los op.
Als mogelijk geef de installatiekosten van de stations los op.

Prijs gasmeters	
Installatie gasmeters	
Aanlegkosten zonder ontwerpkosten, verwijderen oud netwerk en aanleg stations	€

Figure 4.7: Screenshot data gathering spreadsheet part C: costs

The data gathering spreadsheet can be completed by adding the project costs in the final section. As can be seen in Figure 4.7 first and most importantly the **total project costs** are requested (yellow). Additionally also the **material cost, design cost, extra cost for removing old pipes**, and finally **home installation set costs** are requested. These four additional costs are either not likely to be related to the surrounding parameters (material, design, home installation) or not relevant for DH cost predictions (removing old pipes). These costs are requested so that they can be subtracted from the total cost to potentially increase the correlation between the remaining "total cost" and the surrounding parameters that also influence heating networks. The costs are requested only at limited detail because the assumption was made that the costs on this detail are no longer considered as sensitive information by underlying contractors. This assumption was validated by talking to the water and gas utilities and project partners from the consortium.

The material costs are a lot less likely to be related to the surrounding parameters compared to the other project costs. Because of this, it is desirable to subtract these costs from the total project costs, by doing this potential noise that is present in the material cost data is removed from the equation. Furthermore, the material costs are always saved separately in the systems of the market parties and therefore easy to provide.

For the design cost also the assumption is made that they are less likely to be correlated to the surrounding parameters. However, for most companies, these design costs are not saved separately in historical project cost databases. Also, the definition of design cost is harder and therefore more prone to noise since different companies might include different things in this cost component. However, since the design cost is asked separately in the analysis it can be checked whether or not it is desirable to extract these design cost.

The extra cost for removing old pipes is requested, to potentially be subtracted from total costs, for two reasons. First, this cost component is not present when heating networks are constructed since no existing infrastructure exists that needs to be removed. Second, for some projects, the new pipes are not constructed in the same trench as the old pipes which means that twice as much digging is required. If these extra costs would not be removed this will add noise to the cost data. However, as for the design cost, the extra cost for removing old pipes also is not present in cost databases and is therefore also prone to interpretation noise. Also for some water and gas utilities, no information about the removal cost is available at all and the removal costs had to be estimated by the utilities based on removal prices per meter. Even-though the utilities might be capable of reasonably estimating these costs, it should be validated that removing these costs adds value to the modeling performance.

The final cost component, the home installation (gas and water meters), is not present for most projects since these installations are installed and replaced in separate projects. The selection process of the cost components that are considered in the final model is described in section 5.5.

4.2.3. Surrounding parameters from GIS databases

Based on interviews with domain experts and discussions with colleagues from Deltares participating in the research a preliminary overview of potentially important surrounding parameters is constructed. All these surrounding parameters are provided to the ML model as potential inputs. When the model performance is optimized certain surrounding parameters will be included and other surrounding parameters will be neglected. In the end, the best performing models give a nice indication of which surrounding parameters are in fact statistically correlated with the construction cost. More insights and information on the analysis and the "final" selection process of the surrounding parameters can be found in chapter 5.

Table 4.3: Overview surrounding parameters imported from GIS

Category	Surrounding parameters	Unit	Source
Road	Footprint	[m^3]	BRT
	Road coverage	Category	BGT
	Road type	Category	BRT / OpenStreetMap
	Width	[m]	BRT
	Traffic density	Category	-
	Average max speed	[km/h]	OpenStreetMap
	Tunnel in road	[yes / no]	OpenStreetMap
Buildings	Footprint	[m^3]	BAG
	Building year	[#]	BAG
	House type	Category	ARCGIS layer
	Amount of inhabitants	[#]	CBS
	Average distance to road	[m]	Self generated based on BAG
Subsurface	Ground type	Category	Bodemkaart
	Amount of other infra	[#]	-
	Groundwater level	[m]	ARCGIS layer
	Polderpeilen	[m]	DINOloket
	Ground pollution	Category	Data gathering spreadsheet [1]
	Metro lines crossings	[yes/no]	-
	Archeology	[yes / no]	ikaw3
	Explosives	[yes / no]	-
Surface	Private landowners	[yes / no]	-
	Tree density	Category	ANK
	Dykes and waterworks	[yes / no]	BBG
	Area type	Category	BBG
	Railways	[yes/no]	BBG

In Table 4.3 the preliminary overview of the important surrounding parameters identified in the interviews is shown. As can be seen, the parameters are subdivided into four categories to make the overview

[1] Since no database with ground pollution levels was found ground pollution data was gathered using the data gathering spreadsheet

easier to read. This subdivision does, however, not play any role in the statistical analysis. In the last column of the table, the national database from which the surrounding parameter can be imported is shown. As can be seen in the table unfortunately not for all potentially important surrounding parameters a national database that contained the parameter was found.

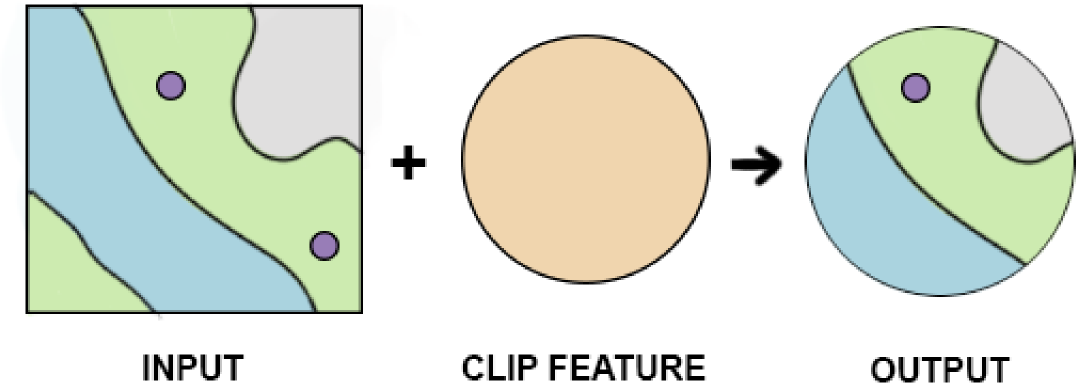


Figure 4.8: Clipping example from ArcGIS website [8]

The national databases that contain the surrounding parameters are too big to use as input for the simulations. For that reason, the databases have to be clipped. Clipping of data is a GIS operation in which a certain data layer is cut into a smaller area to reduce the size of the database. The resulting smaller database can more easily be used to generate the required model inputs. What happens is the following, a smaller area of interest is defined which is called the clipping feature. This clipping feature can be a polygon (regular area) a line or a point. The clipping feature is then compared to the input data layer. Only the area which is present in both the clipping feature and the input data layer is outputted. An example of this can be seen in Figure 4.8.

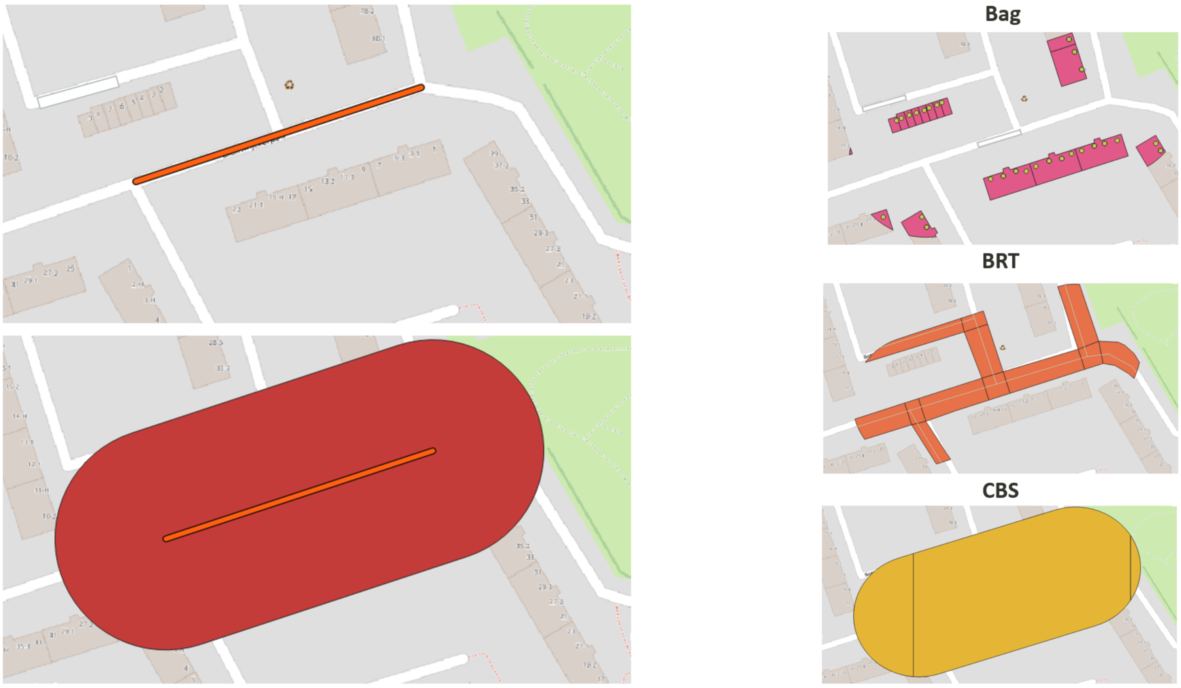


Figure 4.9: Example of clipping GIS databases with buffer around project trace

In this research, the area of interest is the network trace. However, since the network trace is only a line, which for almost every project does not intersect with for example the BAG database containing the information about buildings, this trace can not be directly used as the clipping feature. Instead, a buffer-zone is generated around the network trace, to be used as a clipping feature. This buffer-zone is basically the network trace but then with a very thick line. With this clipping feature, the important parts of all databases shown in Table 4.3 are selected for every single project.

An example of this, selection process, for one very small project is shown in Figure 4.9. As can be seen in the figure the analysis starts with only the trace (top left corner) next a buffer is generated around this trace (bottom left corner). This buffered trace is used to extract the relevant parts of the databases an example is given on the right side for the BAG, BRT, and CBS databases. The information in the resulting relevant parts of the databases is used for the rest of the analysis. For the BAG layer sometimes the clipping needs to be adjusted to remove half buildings. As can be seen in the example in the top right corner of Figure 4.9 sometimes the buffer-zone cuts a building in half, or even less than half. When a building is relatively far away from the trace and only a very small part of this building is actually inside the buffer-zone it is not desired to use the information of this building as input for the model. That is why all the buildings, that do not have their "verblijfobject" point (green points in example) inside the buffer-zone are not considered and also removed from the data layer. For some data layers, some data pre-processing is required before the data can be inputted into the ML model. This can have multiple reasons but the two most common are: the data is categorical value, and a (weighted) average has to be taken because the resulting data layer contains multiple so-called attributes. This data pre-processing is described more elaborately in section 5.1.

An important question that remains in the clipping of the databases is the size of the buffer. As can be expected when the buffer-size changes the surrounding parameters that are considered in the analysis will also change. When the buffer size is too small certain surrounding parameters, like for example the building information from BAG, will not be inside the buffer-zone and therefore not be included in the analyses. However, when the buffer size is too big some surrounding parameters, like for example the presence of water, will be inside the buffer area whilst in reality the water is a couple of streets away. In general surrounding parameters that are further away from the trace are less likely to influence the project cost. Therefore it is very important to choose "the right" buffer size which will include surrounding parameters that are most likely to influence the project and exclude the parameters that do not. Determining the right buffer size leaves room for interpretation since it is not known beforehand which surrounding parameters are in fact influencing the construction cost let alone how close they have to be before they are important. To determine a reasonable buffer size two types of analysis are conducted. First, a sensitivity analysis is conducted to see how much the surrounding parameters change when the buffer size is altered and second a couple of projects are analyzed by hand to see which buffer size seems the most reasonable for those specific projects.

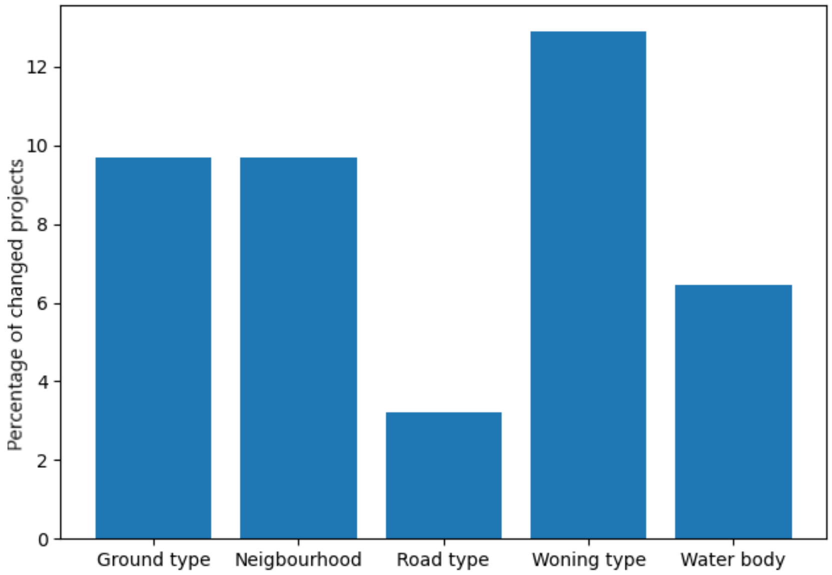


Figure 4.10: For how many (in percentage) of the 31 considered projects did the five considered categorical surrounding parameters not have the same value for the 4 considered buffer values (25m,30m,35m,45m)

Clipping all the databases is a computationally heavy task, meaning that it takes quite long to clip all the databases in multiple different buffer sizes for all the gathered projects. That is why the sensitivity analysis is conducted using only 31 projects and 5 different buffer sizes namely 15m, 25m, 30m, 35m, and 45m. After generating all the clipped data layers for the 5 different buffer sizes it became clear that 15m was too small. Because for 12 out of the 31 projects not a single house was inside the buffer area. So in the rest of the sensitivity analysis, only the other 4 buffer sizes are considered. As mentioned before there are two types of parameters namely categorical and numerical values. In the sensitivity analysis, these parameters were analyzed separately.

First, let us take a look at the categorical values. For the categorical values, it was analyzed for how many of the 31 projects the category of certain surrounding parameters changed over the four different buffer sizes. In other words, does it matter if the buffer-size is 25m, 30m, 35m, or 45m, or does the category value stay the same no matter which buffer size is chosen. The results of this analyses for 5 considered categorical surrounding parameters can be seen in Figure 4.10. It is important to realize that the vertical axes in the figure represent the percentage of projects that changed and not the number of projects that changed. This means that for most projects (approximately 90 %) the category values of the surrounding parameters do not change when the buffer size is varied between 25m and 45m.

$$Difference = \frac{max(x_i) - min(x_i)}{\overline{x_i}} * 100 \tag{4.1}$$

Now let us consider the numerical values. For the numerical values, the percentage difference between the minimum and the maximum value out of the 4 different buffer-size is analyzed. This percentage is calculated using Equation 4.1.

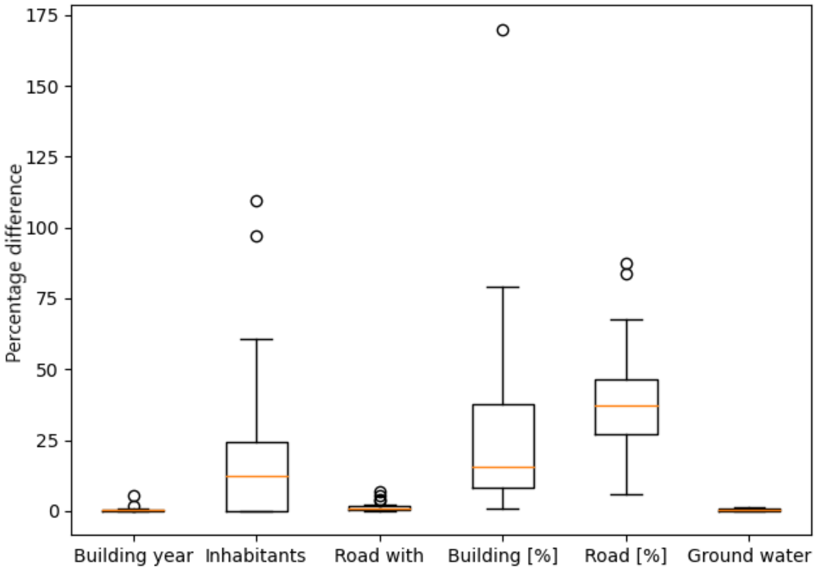


Figure 4.11: What is the difference, in percentage, between the maximum and minimum value of the 6 considered numerical surrounding parameters for the 4 considered buffer sizes (25m,30m,35m,45m)

The results of this analysis are shown in Figure 4.11. As can be seen in the figure for half of the considered surrounding parameters changing the buffer size significantly influences the outcome. For two of these parameters that makes a lot of sense since they are directly dependent on the buffer size. Both building and the road footprint are calculated by dividing respectively the building and road area with the buffer area. The fact that the footprints change that much when the buffer size changes is therefore logical but it does not give any indication on which buffer size is the best.

The final surrounding parameter that seems to change significantly when the buffer size is changed is the number of inhabitants. However, when taking a closer look at the projects where the variation is really big the conclusion was drawn that the percentage difference is so big because those projects have a lot more inhabitants than the average number of inhabitants. So when you would look at the percentage difference between the 4 buffer sizes and divide it by the average of these four different buffer sizes instead of the average

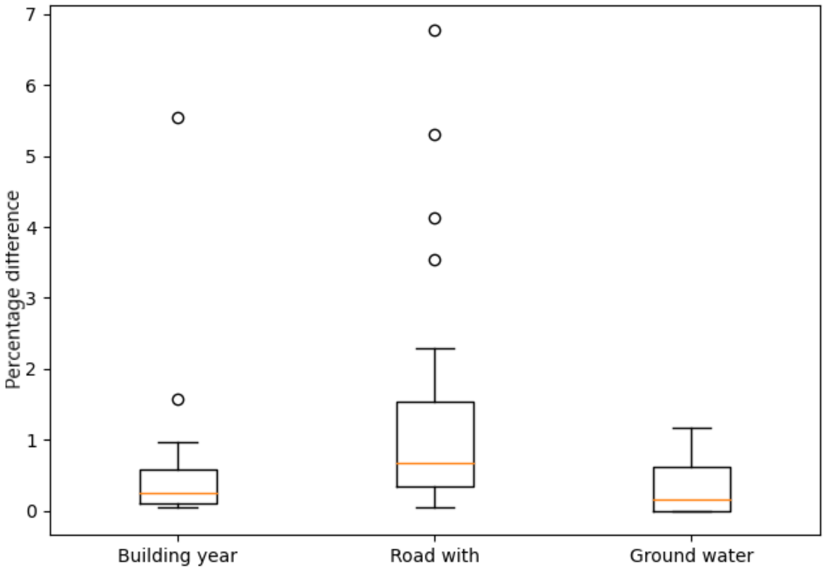


Figure 4.12: What is the difference, in percentage, between the maximum and minimum value, of the 3 numerical surrounding parameters with the smallest differences, for the 4 considered buffer sizes (25m,30m,35m,45m)

of all projects the percentage differences for these projects are a lot smaller. When looking by hand at the differences in the number of inhabitants of different projects the conclusion was drawn that the amount of inhabitants indeed is dependent on the buffer size but not as strong as the boxplot in Figure 4.11 would suggest. The other three surrounding parameters don't vary that much when the buffer size is changed. In the zoomed-in plot in Figure 4.12 it can clearly be seen that the percentages are very low.

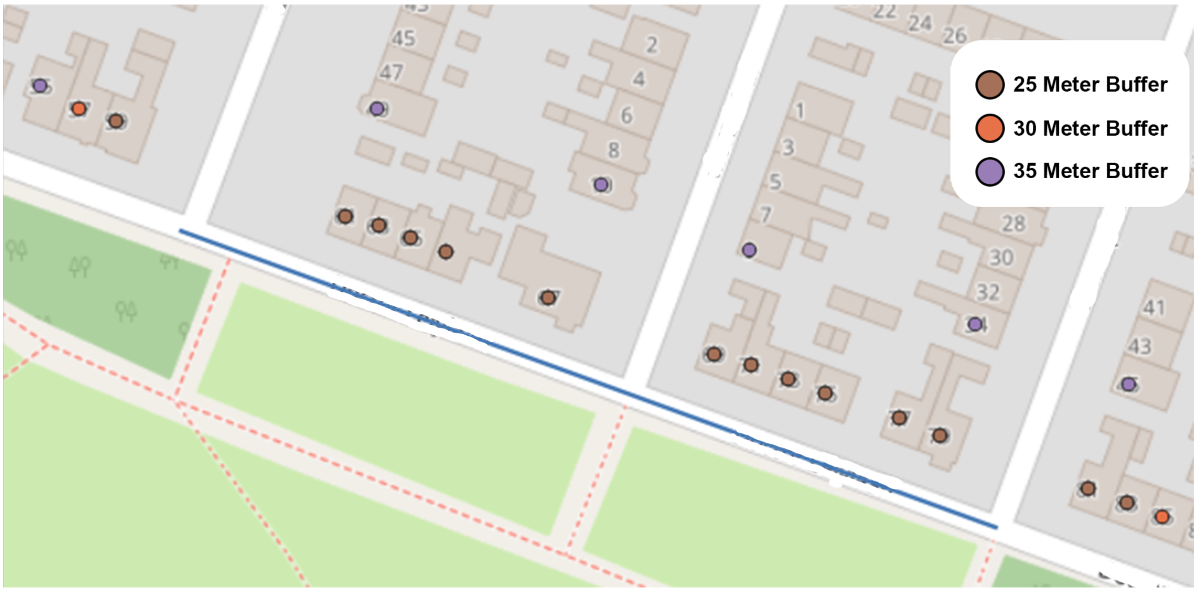


Figure 4.13: Example chosen buildings with different bufferzones

Based on the sensitivity analysis the conclusion can be drawn that for most surrounding parameters that were checked the size of the buffer does not influence the results significantly. However, since this analysis was only based on a segment of all the projects used in the final model and also not all the surrounding parameters were analyzed, it is decided to do another sensitivity analysis after the final model is developed. In this analysis, all projects are considered but only the surrounding parameters that made it to the final model are checked. The results of this analysis are presented in the discussion. Because of this chosen strategy and because most surrounding parameters are not influenced that much choosing a buffer size by hand by studying some examples projects was deemed a reasonable solution.

After studying a couple of example projects it was concluded that 30m was the most logical buffer size. In the example shown in Figure 4.14 it can be seen that the 30m buffer did the best in selecting houses that were facing the trace. Of course, not all examples were as clear as this one when looking at the project in Figure 4.13 for example it can be seen that for this project 25m would have been a better buffer size. However, in this project, 30m is still reasonable where ass 35 is clearly too big. After analyzing 10 example projects it was concluded that 30 meters maybe was not the best solution all the time but it did provide a reasonable buffer zone for all of them and is therefore the best overall buffer size.

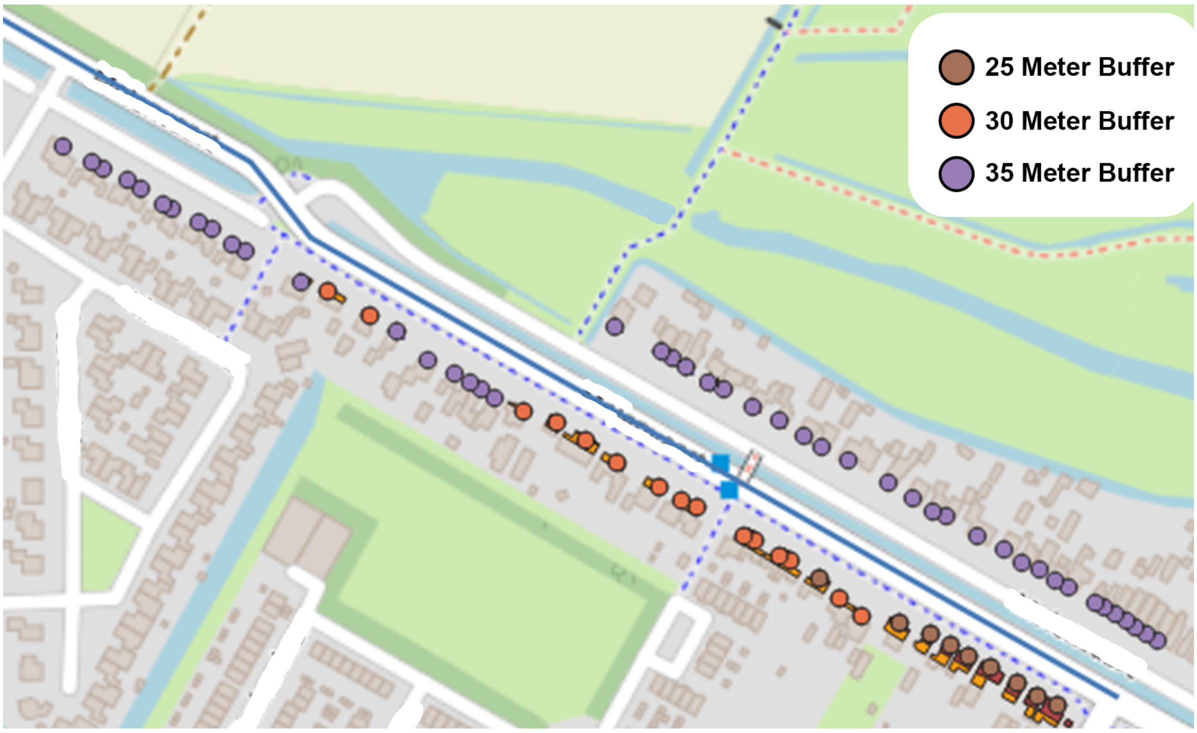


Figure 4.14: Example 2 chosen buildings with different bufferzones

5 Developing the Model

In chapter 2 it was concluded that linear regression was the most suited ML algorithm for this research. However, applying linear regression to model costs can be done in a lot of different ways. This chapter describes how a linear regression model for construction costs prediction can be developed using the input table filled with the spreadsheet and GIS data described in chapter 4. The chapter is structured in the following way. In section 5.1 it is explained how the original input table is altered and checked so that it can directly be used as input for a ML model. Afterward, different strategies to develop and improve a regression model are discussed. First, in section 5.2 the main scoring criteria and the baseline performance are presented. Then in section 5.3 it is discussed how the right input parameters can be chosen from the total input table, also it is discussed how the modeling performance can be improved further by altering the input parameters in smart ways. In section 5.4, it is explained how random (dummy) project data is generated for the dummy analysis. This dummy analysis is conducted to see how well ML can reconstruct a predefined and therefore known cost function and to check the influence of the number of projects on the modeling performance.

With the modeling strategies described in this chapter, five different types of models are generated: models based on water data (water model), models based on gas data (gas model), models using both water and gas data to predict gas (combined model gas), models using water and gas data to predict water (combined model water) and finally a model using the generated dummy data (dummy model). The combined models are generated to check whether a model using both water and gas data for training has a better performance for predicting water and gas costs than the separate water and gas models that were trained only using their own historical project data. When this is the case it more likely that using the water and gas modeling information is also of added value for DH modeling, which is the main goal of the research.

Even though all the modeling strategies that are described are considered for all models. It is important to note that not all the things described in this chapter are also implemented in the final models. Since not all the mentioned things did improve the performance. Also, it is possible that a certain approach did increase the performance of the gas model when it did not increase the water model performance. Meaning that the resulting models for the different data sets could be different. The resulting models and their performance are presented in chapter 6.

5.1. Data pre-processing

Before the raw data from the data gathering spreadsheet and the selected (clipped) areas of the national GIS databases can be inputted into a ML model first some alterations and checks are required. In this section, this data prepossessing is discussed. First, in subsection 5.1.1 the alterations to the spreadsheet data are discussed. Secondly, the adaptions to the GIS data are presented in subsection 5.1.2. Finally, the process of checking the resulting input table for outliers and missing data points is elaborated on in subsection 5.1.3.

5.1.1. Spreadsheet data alterations

As explained in subsection 4.2.2, the data gathering spreadsheet consist of three parts. In this subsection, the data alternations are described per part. As also discussed in the data gathering spreadsheet subsection, not all the data that is requested in the spreadsheet was available or deemed relevant by the water and gas utilities. In this section, only the inputs that were provided by a significant amount of utility companies are discussed.

This means that all the inputs that are part of the data gathering spreadsheets see Figure 4.5, Figure 4.6 and, Figure 4.7, but are not discussed here are not used as inputs for the ML model. Because these inputs were either not available or not deemed relevant.

The first part of the spreadsheet contained the project's characteristics. The project characteristics that are used are: **start and end date** (both planned and real), the **amount of other infrastructures** involved in the project with corresponding meters, and the **ground pollution** category also with corresponding meters.

Since dates themselves can not be inputted into a ML model it was decided to use the **total amount of days** between the start and end date as input for the model for both the planned and real dates.

For the **other infrastructures**, the amount of other infrastructures that participated in the project was chosen as an input. This means that it was neglected whether a certain infrastructure participated for the entire project or only a part. This decision was made because detailed information on the number of meters that were realized together compared to the number of meters that were done alone was not available for most companies.

For the **pollution category** first, the pollution category with the highest amount of meters was selected. The resulting pollution category than was changed to a numerical value using the logic shown in table Table 5.1. It was chosen to merge the no pollution and light pollution groups because it turned out that construction projects are always realized using so-called basis hygiene measures. This means that projects with no safety measures for pollution do not exist. The resulting pollution input can be used both as a categorical and numerical input for the ML model according to the logic in Table 2.5.

Table 5.1: Pollution category to numerical input

Pollution name	**Model input**
No pollution	0
Light pollution	0
Medium pollution	1
Severe pollution	2

In the second part of the spreadsheet Network characteristics were requested. The network characteristics that were used in the analysis are: **Total amount of connections** in project area, Amount of **replaced connections**, **soil replacement** (yes/no), **Dewatering** (yes/no), **material and diameter** with corresponding meters and **construction method** with corresponding meters. The first four spreadsheet inputs can be used directly in the model. However, the information about the pipes and the construction methods have to be altered before they can be used as inputs for a linear regression model.

The alteration of the **construction methods** is the most straightforward. The seven different construction methods all get their own category which is filled with the number of meters that a certain construction method is used in that specific project. The seven different construction methods are: **Open trench, Drilling, Sinker, Pressings, above the ground, in an existing pipe (ILT) and other (unknown)**.

For the **material** originally two different approaches are chosen so that the best approach can be identified in the model training and performance evaluation described in the last three subsections of this chapter. In the first approach, the **average diameter** and the most occurring material type (based on length) are used as inputs for the model. The different **material types** for water and gas are: PVC, PE, Steal (ST), gray cast iron (GIJ), asbestos cement (AC), and other. In the second approach, the material types and diameters are combined in 9 categories of three different material types and three different diameter ranges. This second approach is chosen chosen to validate weather certain combinations of the diameter and the material type have a bigger impact on the construction costs than others. The chosen categories with corresponding diameter ranges are shown in Table 5.2.

The final output of this part of the spreadsheet is the **total length** of the project. The total length of the project is calculated by adding up all the lengths in both the materials and the construction methods table and taking the average of both lengths. In theory, the total length of all the materials should be the same as the total length of the construction methods. However, since for some utilities this information is acquired from a different location in the project database sometimes there is a difference between these total lengths. The assumption was made that both total lengths are equally suited for calculating the final total length and therefore the average is used.

Table 5.2: Second approach for choosing material and diameter categories that are used as model inputs

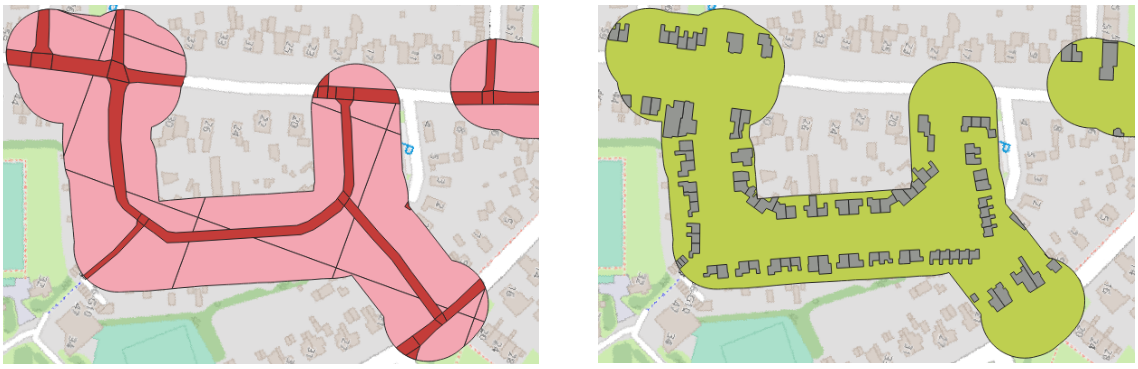
	PVC			**PE**			**Other**		
	Small	Medium	Big	Small	Medium	Big	Small	Medium	Big
Min diameter [mm]	-	76	126	-	76	126	-	76	126
Max diameter [mm]	75	125	-	75	125	-	75	125	-

The third and final part of the spreadsheet contains the cost data. The cost data is the so-called dependent variable that needs to be predicted by the generated ML model. The chosen dependent variable in this analysis however is not the total project cost but the construction cost per meter. To create this construction cost per meter the construction cost has to be divided by the total length of the project. The construction cost itself can be created in different ways from the cost data from the spreadsheet. As already stated in subsection 4.2.2, not just the total project cost is available but also the material cost, cost for removing old pipes, design cost, and home installation cost (which is 0 for most projects). Subtracting these above-mentioned costs from the total project cost might lead to a construction cost that is more correlated with surrounding parameters and therefore is easier to model. However, this needs to be validated when the different models are trained and their performance is compared. That is why the different costs are all provided as potential labels for the data that is inputted into the model. The process of choosing the "best" dependent variable is described in section 5.5.

5.1.2. GIS data alterations

The clipped GIS data layers containing the surrounding parameters for all the projects can not be inputted straight into a ML model. First, it is important to realize that most GIS layers, also called shapefiles, consist of multiple attributes. This for example means that the CBS shapefile, which contains data on the number of inhabitants per hectare, does not provide one number for the entire project. Instead, every attribute of this shapefile has its own value for the amount of inhabitants. How many attributes there are in a project depends on the size of the buffer zone of that project and how the attributes are distributed in the original layer of the entire database. This distribution of attributes is different for different databases and is therefore also different for all the shapefiles of the same project.

An example of this is presented in figure Figure 5.1. In Figure 5.1a the CBS shapefile containing the inhabitants per hectare (pink) and the BRT shapefile containing information about the roads (red) are shown. As can be seen in the picture the CBS shapefile is divided into rectangles, these rectangles are precisely one hectare and cover the entire country (if there are no inhabitants there is no rectangle). The road attributes are separated by intersections and also cover the entire dutch road network. In Figure 5.1b the average groundwater level shapefile is shown (green) together with the BAG shapefile (gray). It can be seen that the attributes of the groundwater level data layer, for this project, are a lot bigger than for example the inhabitants' layer. As could be expected the BAG data layer has one attribute for every house in the Netherlands.



(a) GIS shapefiles CBS (inhabitants) and BRT (road info)

(b) GIS shapefiles AVG Groundwater level and BAG (buildings)

Figure 5.1: Examples of different shapefiles for same project

Every attribute in a shapefile can contain multiple datapoints. The BAG shapefile for example contains data about among other things, the building year, the building surface, and the building status. Before the data can be inputted first the right data point for every attribute has to be selected. Then all these data points have to be merged into one data point that represents the entire project. The merging of all data points into one project data point is different for different surrounding parameters and is discussed below.

First, it is important to splits all the surrounding parameters into two groups the surrounding parameters that have a numerical value and the surrounding parameters that have a categorical value. The surrounding parameters with numerical values are easier merged into one data point that can go into the model and are therefore discussed first.

The most straightforward way to merge the numerical data points is by taking the average (mean) of all attributes. This is done for the Building year surrounding parameter. However, instead of taking the average building year as model input, this average building year is subtracted from 2020 to get the average amount of **Years ago** all the buildings in the project area built. It is chosen to input the amount years ago instead of the building year because the years ago possess the same information but by using smaller numbers. For a linear regression model, it is desirable that all variables have a similar order of magnitude. Besides inputting the amount of years ago also an input was created to represent the diversity of the buildings in a project area. This input was created because it is expected that a higher diversity of building years in a project area means that more different building projects occurred in a neighborhood, which is likely to create more chaos (in the subsurface) and could therefore lead to higher costs. This diversity input is generated by subtracting the mean of all building years from the median of all building years in the project area. The resulting **Building year diff med and mean** is used as model input.

However sometimes simply taking the average gives a twisted representation of the entire project. When considering the inhabitants' data layer for example it can be seen in Figure 5.1a that some attributes are a lot smaller than others. Simply taking the "normal" average of all attributes would give these small attributes more "power" than they deserve. That is why for the **inhabitants**, the **average groundwater level**, the **polderpeil peilgebied**, the **percentage trees** and the **trefkans archelogy** the weighted average based on area is taken instead. When doing this it was discovered that the polderpeil data layer did not cover the entire country and therefore should be used with care. A final input for which the weighted average is taken, but this time based on length instead of area, is the average speed limit (**Average maxspeed**).

There are also some surrounding parameters for which the data is not saved in one of the attributes but is dependent on the shape/area of these attributes. This is the case for the **footprint of the building** , the **footprint of the road**, the **width of the road** and the **distance of the houses to the road**. To calculate the two footprints the total area of the road and the buildings are divided by the total buffer area. For the road width calculation, the area of the road is divided by the length of the road. Finally, the distance of the houses to the street is calculated for every house using the Geopandas distance function [3]. Then the average of all these distances is calculated to serve as input for the model.

Now let us consider the categorical surrounding parameters. The categorical surrounding parameters can again be subdivided into two groups. The first group indicates the presence of something in the project area and the second group is more descriptive. The four surrounding parameters in first category are: **water**, **railroad**, **cley** and **tunnel**. These four surrounding parameters are 1 if the project area contains water, railroad, cley or a tunnel and or 0 when no water, railroad, cley, tunnel is present in the project area.

Table 5.3: Categorical descriptive surrounding parameters overview

House type	Neighborhood type	Road type	Road coverage
Appartment	**Residential Area**	Motorway	Closed hard coverage
Detached house	Industry	Highway	Half hard coverage
Terraced house	Forrest	**Street**	Open hard coverage
Corner house	Park	**Local road**	Sand
Semidetached house	Sports	Regional road	
	Retail	Parking place	
	and more less relevant ones		

The four surrounding parameters in the second category are: **house type**, **neighborhood type**, **road type** and **road coverage**. These surrounding parameters can have a couple of different "values" as can be seen in Table 5.3. Of course, it is also possible that different values of the same surrounding parameter are present in a project.

To solve this "problem" two different strategies were applied. The first strategy is applied for House type, Neighborhood type, and Road type. For this first strategy, when multiple values occur in the project area, the value which has the highest total area is considered as the surrounding parameters of that project. Because the total amount of different values is quite high it is chosen not to import all of them into the regression model. Based on an evaluation of all the project data the most common values of the surrounding parameters are chosen. When those values are the surrounding parameters of a certain project a 1 for that surrounding parameter is inputted in the model. The chosen most common surrounding parameters are bald in the table. Besides the bold categories shown in the table also the category other is used as input, for the three surrounding parameters. The other input parameter is 1 when the value with the biggest area was not one of the chosen (bold) values and 0 otherwise.

The second strategy that was used to differentiate when different values were present in a categorical surrounding parameter, was only applied for the road coverage. Instead of choosing the most common road coverage the percentages (based on area) of all the four different road coverages (**percentage asfalt**, **percentage half verhard**, **percentage open verharding**, **percentage zand**) were used as model inputs.

5.1.3. Check resulting input table

The checking of the input table consists of two things. First, it should be checked whether the input table is complete, meaning that for every project all the required inputs described in section 5.1 should be available. Secondly, the input table should be checked for outliers. When outliers are found it is important to identify whether such an outlier is a faulty data point and should therefore be removed or corrected or that the outlier just describes a real extreme situation that should be included in the model. When checking for outliers, the most extreme data points (projects) for both the water and gas databases are checked by looking at the data gathering spreadsheets again by hand, to see if these extreme values are likely to be extreme cases or faulty data. For non of these extreme projects, a fault was detected, and therefore no data points were removed from the input table.

When looking at the total input table of the water and gas projects it was analyzed that only 5 projects did not contain all the required inputs. For three of these projects, the average groundwater level was missing. When looking into this issue it was concluded that this was because the projects were outside of the "scope" of the average groundwater data layer. Two of the projects are on Texel and one at the Maasvlakte, both these areas are not included in the database. It was decided that these three projects were used as input for the model except when the groundwater level was considered as one of the inputs. The other two projects that were missing inputs did not contain any information from the BAG. This was because for these two projects no buildings were present in the buffer zone. Because this analysis focuses on construction cost in the built environment these projects were deemed not relevant for the research and therefore removed from the database.

5.2. Scoring criteria and baseline

When training and optimizing a linear regression model, or any other ML model, it is important to first identify a scoring criteria and a minimum achievable performance level. This minimum achievable performance level, also called the baseline, represents the currently available models. When the newly developed model performs less than this baseline the new model does not provide any added value. When this is the case either the model should be developed further or the conclusion should be drawn that the modeling approach is not sensible. Without a scoring criteria, it is not possible to check whether any changes to the model actually improve the model performance, and also it is not possible to check if the model outperforms the baseline. As presented in subsection 2.2.1 there are multiple scoring criteria available for regression models. However, in this research, it is chosen to choose a single main scoring criteria to evaluate the model performance in the optimization process. Using multiple scoring criteria when optimizing a ML is more challenging since some alterations may improve certain scoring criteria when other scoring criteria decrease. Every time this happens the most desired situation has to be chosen, this makes the optimization process a lot harder to automate which is desired in this research.

In this research is it chosen to use the most common scoring criteria the coefficient of determination, also known as the $\boldsymbol{R^2}$ score. This scoring criteria is chosen both because it is common and because it is straightforward to compare the model performance of different models. When using MSE, another common scoring criteria, the model performance is expressed using the size of the error in the predictions. The downside of this scoring criteria is that when datasets have bigger errors the model seems to perform less compared to datasets with smaller errors. This phenomenon (some datasets are harder to predict than others), which can affect any scoring criteria, is another reason why it is desirable to compare the model performance to a baseline (using the same database) instead of comparing the model performance to models trained using different datasets.

As also described in subsection 2.2.1 the $\boldsymbol{R^2}$ scoring criteria can be used in two different ways. The goodness-of-fit can be calculated using training data (y_{train},x_{train}) and the prediction accuracy can be computed using, unseen, test data (y_{test},x_{test}). To score the performance of the model the **prediction accuracy** is used. It is chosen to use the prediction accuracy because it is more important to be able to predict costs for new unseen projects than to predict the cost of already realized projects.

The baseline model chosen in this research is the linear regression model using only the diameter as input to predict the construction cost per meter. This baseline is chosen because most currently available models only use the diameter for cost predictions and in The Netherlands, there is no model available that makes more detailed construction cost predictions. The R^2 scores of the baseline water, gas, and combined models are presented in Table 5.4. The performance of the model is dependent on the division of the total data set into a training and test dataset. This can clearly be seen in the table were k-fold splitting (k-fold splitting explained in subsection A.0.1) is used to do this splitting five times in a different way.

Table 5.4: Baseline R^2 scores for the different kind of models for the different splits resulting from k-fold splitting

Model type	**Baseline R^2 score**						
(data points train / data points test)	**Average**	split 1	split 2	split 3	split 4	split 5	split 6
Water (57/11)	-0.17 (0)[1]	-0.79	0.36	0.20	0.021	-0.46	-0.36
Gas (42/10)	0.16	0.046	0.32	0.35	0.12	-0.039	-
Combined Water (109/11)	-0.26 (0) [1]	-1.01	0.57	0.18	0.071	-0.59	-0.61
Combined Gas (110/10)	0.15	0.041	0.28	0.27	0.13	0.039	-

The scores of the five models differ quite a lot for the four different model types. It is chosen to take the **average of the n-models** resulting from k-fold splitting as **main scoring criteria** for this research. This decision is made to minimize the effect the splitting of the data has on the model performance. It should be noted that even when applying this approach the random splitting of the test and train data influences the model performance. To make sure that the results of the analysis are reproducible a random seed is used when the train and test data is split.

As can be seen in the table the baseline model performances of the Water, Gas, Combined Water, and Combined Gas, are $-0.17, 0.16, -0.26$, and 0.15 respectively. For both the water model a negative R^2 value is found. At first, this might seem impossible since a squared number can not be negative. However, when taking another look at Equation 2.13a, it can be seen that R^2 is not actually a value R squared, and can therefore be negative. A negative R^2 means that the model performs worse than just taking the average (construction cost) for every prediction. This of course is not very common and indicates that your problem is useless. For that reason, it is chosen to use a $\boldsymbol{R^2}$ **of 0** as baseline performance for both the **water models**.

Table 5.5: Baseline MSE scores for the different kind of models for the different splits resulting from k-fold splitting

Model type	**MSE test score**						
(data points train / data points test)	**AVG**	split 1	split 2	split 3	split 4	split 5	split 6
Baseline Gas (42/10)	67844	154854	72876	36494	41385	33622	
Combined gas (110/10)	69271	155680	77258	41130	41172	31116	
Baseline water (57/11)	29602	8202	68574	27440	5604	40954	26837
Combined water (109/11)	27717	9262	46508	28039	6129	44560	31807

Even though the model is optimized using a single main scoring criteria, $\boldsymbol{R^2}$, it is chosen to also use another scoring criteria to validate the final model results. This is done because a certain model can achieve a very high score on a certain scoring criteria based on random chance. When this same model is scored using different scoring criteria these other scoring criteria can indicate that in fact, the model does not perform

[1] For both water models it is chosen to set the baseline model performance at 0 instead of the negative model performance of the water models only using the average diameter to make predictions

properly. To minimize this risk, another common scoring criteria, **MSE**, is used to score both the baseline models and the final best performing models. This validation is used to check that indeed the best performing models perform significantly better than the baseline models. The baseline model MSE scores are presented in Table 5.5 and the MSE scores of the best performing models are discussed in subsection 7.5.2.

5.3. Choosing input parameters

When all the input parameters from the spreadsheet and GIS databases are combined every project has 47 potential inputs. This is less than is sounds because 13 out of these 45 are related to the material and the diameter (9 categories from table Table 5.2, average diameter and 3x material). Also from the input parameters that together represent one categorical input value, for example, the road type inputs Street, Local road, and other road, always one of these parameters should be considered as the base scenario and therefore not be used as input for the regression model. So if a ML model was developed using only the road type as input instead of using all the three categorical input variables only two should be used. Choosing the default category in some is straightforward and sometimes a random default category has to be chosen. For the performance of the model, the choice of the default category is not important. The principle of always using one variable less by choosing a default value is called the **dummy variable trap**. This is because the multiple variables describing the same categorical input are called dummy variables.

Still, the amount of remaining inputs is too high and only a subset of the total input parameters should be inputted into the model. As a rule of thumb it desirable that the ratio between the amount of projects and the input parameters is between 15:1 and 20:1 but it should definitely not be smaller than 5:1 [33]. This means that for example for the gas model which only has 52 projects a maximum of 10 inputs is allowed. In this section, first, three different strategies for selecting the "best" subset of input parameters are discussed, and afterward, potential alterations of input parameters are discussed.

5.3.1. Three approaches for choosing inputs

When selecting the "best" subset of input parameters out of the total potential input table three general approaches can be used. None of these approaches is necessarily better than the others and most of the time it is valuable to try all the methods and see which method gives the best results. The three methods are[33]:

- *Forward regression:* Start with the most promising parameter and every iteration add the most promising parameter that is not yet included. Stop when adding more parameters decreases the model performance.

- *Backward regression:* Start with all parameters and every iteration remove the parameter which seems to perform the worst. Stop when removing parameters starts to decrease the performance instead of increasing it.

- *Stepwise regression:* Again start with the most promising parameter and add the most promising leftover parameter every iteration. However, this time, every iteration it should be checked whether all before included parameters still provide added value. If not remove these parameters before the next iteration. Stop when both removing or adding new parameters does not improve the modeling performance.

The question that remains is how to identify the worst-performing parameter when removing parameters and the parameter with the most potential when adding parameters. Ways to identifying these parameters for both scenarios are presented in Table 5.6

Table 5.6: Ways to identify potential parameters to add or remove from model

Identifying high potential parameters to add	**Identifying poorly performing parameters to remove**
Look at correlation between dependent variable (y) and considered input parameters (x). High correlation means higher potential	Look at the t-statistic of all the parameters the lower the t-statistic the higher the significance of this parameter
-	-
Calculate the Tolerance value of a potential input the higher the Tolerance value the higher the potential. See subsection A.0.3 for a short explanation of the Tolerance value calculation process.	Apply L^1 regularization, See subsection A.0.4 for a short explanation of the different types of regularization
	-
	Compare the different estimations of the regression parameter of a certain input parameters for the different model if the estimates are far apart the parameter is likely to be insignificant. (when using k-fold splitting)
	-
	Look at the correlations between different inputs. When to inputs are strongly correlated they are not likely to both have an separate added value to the model.

5.3.2. Altering input parameters to increase performance

Besides choosing the rights inputs to improve the performance of a ML model, it is also possible to alter input parameters to improve the model performance. Before the different ways on how to alter input data effectively are discussed it is important to note that it is common practice to first try the model performance with the standard data. It is important to know how well the models perform in this simple configuration before different alterations can be tried. This is because it might be that the simple input data already performs satisfactory, or that the normal input data outperforms the adapted data.

The three ways data can be altered to increase the modeling performance are:

- Interaction terms, see subsection A.0.5 for a short explanation.
- Non-linear transformations, see subsection A.0.6 for a short explanation.
- Change the chosen dummy variables (categories). This can be done by regrouping some of the variables together to create less inputs or by splitting some variables to create more inputs.
- Combine multiple input parameters to create new inputs and/or categories. An example of this would be to label a project with an old building year and a lot of water into the category canal house.

5.4. Dummy projects

Besides using the actual project data from the data gathering spreadsheet and GIS databases as inputs for the ML models also dummy data and used in the 'dummy analysis'. This dummy analysis is conducted to check two things. First, the dummy analysis can be used to validate the modeling approach. In this part of the dummy analysis, it is checked how well a multivariate linear regression model can reconstruct a known cost function. The complexity of the cost function, and the amount of noise that is added, are increased to see if a ML model can filter the noise and identify complex cost functions. This validation is important because it is likely that the real unknown, cost function is rather complex and that the data used for predicting this function has significant noise. The chosen ML model (linear regression), should therefore be able to deal with these challenging circumstances.

The second reason for conducting a dummy analysis is to get a feeling for the amount of input data that is required to make accurate predictions. This is important because the amount of data points (52 Gas, 68 Water) is relatively low for training a ML model. Therefore different amounts of dummy projects are generated for the different complexities and noise levels.

5.4.1. How is the dummy data generated?

The generated dummy data is based on the actual received project data. However, only the independent input parameters of the dummy analysis are based on the received data. The dependent variable (y), cost per meter, is calculated using a chosen cost function. The different cost functions that are used to calculate the cost per meter for all dummy projects are presented in subsection 5.4.2. It is chosen to base the independent input variables (X) of the dummy project on the actual data for two reasons. First, the dummy data has to be as realistic as possible so basing it on real data points assures that the dummy data is similar to real data. This is desirable because potential correlations between different input parameters can improve or decrease the model performance, these correlations should therefore also be present in the dummy data. Second, automatically generating dummy projects based on real projects is a lot faster than producing all dummy projects by hand.

The first step in generating n dummy projects is selecting n random projects from the total database containing the 52 gas and 68 water projects. The n projects are selected in such a way that a certain project may be used 3 times whereas another project is not used at all. To make sure that the dummy projects differ from the real projects normal distributed noise is added to all the numerical input parameters. All the categorical input parameters of the projects are kept the same. The normal distributed noise has the following characteristics: $\mu_x = 0$ and $\sigma_x = 0.1 * x_{max}$. After the noise is added values that are supposed to be integers (Planned time, other infrastructures, Pollution category, Building year, etc) are rounded to the closest integer and finally, all the negative input values are replaced by 0. The n resulting projects, which have no cost data yet, have similar characteristics with the gathered water and gas projects but are not the same.

5.4.2. Added noise and amount of dummy projects

The input table, containing n dummy projects, described in the previous section can not directly be used in the dummy analysis. First, cost data and noise has to be added to the projects. As described above it is desirable to have dummy projects with different noise levels and different cost functions. To identify the effects of increasing the noise or complexity of the cost function to the model performance. It was chosen to use three different cost functions with different complexities (easy, normal, hard) and also three different levels of noise (easy, normal, hard). See Equation 5.1, Equation 5.2, and Equation 5.3, for the three different cost functions and Table 5.7 for the noise levels.

For the noise level of the dependent variable (y) both the sigma as the resulting average noise added are presented. For all the independent variables (X) the sigma of the normally distributed noise is different and is therefore not shown in the table. The sigma of independent variables is a certain percentage, given in the table, multiplied by the mean value of the independent variable for which sigma is calculated. It is important to note that the noise added to the independent variables (X) is added after the independent variables are entered in the cost functions. This means that the resulting data does not only contain noise in the dependent variable, cost per meter, but also noise is added to the input variables. This is done to mimic potential mistakes in the data gathering process, for example, the saving of project data by the companies, and errors in the GIS databases but also mistakes when filling the data gathering spreadsheets.

Table 5.7: Noise levels dummy projects

Noise category	Sigma y	avg percentage noise y	sigma X
Little	35	10%	10%
Normal	50	19%	15%
Hard	70	64%	30%

When varying the complexity of the cost functions the noise level is kept constant on the easy level and when varying the noise levels the cost function is also kept the same (easy level). The 6 different dummy simulations are all conducted using 50 datapoints and 150 data points to identify which noise and cost function complexity levels benefit from adding more datapoints. This means that in total 12 ($6 * 2$) dummy analysis can be conducted. The results of this analyses are presented in section 6.4.

$$\begin{aligned} y_{easy} =& 150 + 0.2 * \text{AVG\_diameter}^{1.3} - 4 * Ln(1 + Length\_open\_sleuf) \\ &+ 0.3 * year\_ago + 1.5 * Inhabitants + 110 * Water\_presence \end{aligned} \tag{5.1}$$

$$
\begin{aligned}
y_{normal} =& 150 + 0.2 * \mathit{AVG\_diameter}^{1.3} - 4 * \mathit{Ln}(1 + \mathit{Length\_open\_sleuf}) \\
& + 0.1 * \mathit{year\_ago}^{1.7} + 30 * \mathit{Inhabitants}^{0}.5 + 3.5 * \mathit{Road\_overig} \\
& * \mathit{AVG\_maxspeed} - 29 * \mathit{Road\_overig} + 1.5 * \mathit{AVG\_maxspeed}
\end{aligned} \tag{5.2}
$$

$$
\begin{aligned}
y_{hard} =& 50 + 0.2 * \mathit{AVG\_diameter}^{1.3} - 4 * \mathit{Ln}(1 + \mathit{Length\_open\_sleuf}) + 0.1 * \mathit{year\_ago}^{1.7} \\
& + 30 * \mathit{Inhabitants}^{0.5} + 3.5 * \mathit{Road\_overig} * \mathit{AVG\_maxspeed} - 29 * \mathit{Road\_overig} \\
& + 1.5 * \mathit{AVG\_maxspeed} + \frac{800}{\mathit{Percentage\_trees}} + \mathit{Width\_road}^{1.9}
\end{aligned} \tag{5.3}
$$

5.5. Develop and improve the model

In this chapter multiple tools are described which can be used to develop and improve a linear regression glsml model. As mentioned in the introduction of this chapter these tools are used to generate five different types of models (Water, Gas, Combined Gas, Combined Water, Dummy). How these five models are developed is described in this subsection.

The first step in the process of training the model was choosing which costs were included in the dependent variable (y). As mentioned in subsection 5.1.1 five different cost are available namely: Total cost, material cost, design cost, removal of old pipes cost and home installation set cost. The home installation cost was 0 for most projects and was therefore not considered. The total cost should always be included in the price per meter, but the other cost can be subtracted. There are two ways this can be done, either the cost components in the dependent variable are chosen beforehand and a model is developed which can predict this specific dependent variable as good as possible or during the improvement process, the different dependent variables are all considered to see for which dependent variable the model can make the best predictions. In this analysis, it was chosen to use the first method. Only the material cost is subtracted from the total cost to create the dependent variable (y) which is used for training all models in this study.

This method is chosen because the assumption was made that the design cost and the removal cost are not accurate and will therefore add extra noise to the model. Because of this, they are not likely to add value but the process of improving the models is significantly more complicated when the dependent variable that needs to be predicted changes during the model development phase. The assumption that the removal cost and the design cost are not accurate is based on interviews with the experts providing the historical cost data. For most utility companies these costs were not directly available and assumptions had to be made to estimate these costs.

The reason why the material cost is subtracted instead of just using the total cost is twofold. Firstly, the material cost is very easy to predict for new projects, using simple price lists, and therefore it is not required to include these costs in the model. Second, the prices of materials are likely to change over time and it can very well be that different prices (due to different profit margins) are used for the same materials by different companies. When the material costs are subtracted from the total costs these uncertainties are removed from the equation.

To validate that subtracting the design and removal costs from the total cost decreases model performance the baseline model (only including diameter) performances of the water and gas model for the different options of the dependent variable are compared and shown in Table 5.8. In this table, it can indeed be seen that subtracting the design and removal cost, at least for the baseline model, decreases the modeling performance.

Table 5.8: Baseline model performances for different choice of dependent variable (y)

	Total - Material	**Total - Material - Design**	**Total - Material - Design - Removal**
R2 test Water	-0.17	-0.28	-0.69
R2 test Gas	0.16	0.14	0.07

Now that the independent inputs (X) are altered and checked, see section 5.1, the main scoring criteria and baseline models are picked, see section 5.2, and the dependent variable (y) is chosen the ML model can be developed. In this study it was chosen to use the Python package **Statsmodels** to to train the linear regression models [16]. For training regularized models, see subsection A.0.4 for explanation regularization, a specific module of the linear regression package of Statsmodels is used named **fit_regularized** [10].

When improving the modeling performance of a linear regression model two things have to be optimized. Firstly, the best possible set of input parameters has to be chosen from the total input table, containing the 47 potential inputs, using the three approaches for choosing inputs described in section 5.3. Keep in mind that the "best" input set can also include (non-linear) transformations of the original inputs as described in subsection A.0.6. Secondly, the best modeling settings have to be identified. The modeling settings that are optimized in this research are the regularization terms ($L1_{wt}$ and λ), the $bagging_{boolean}$ (whether or not bagging is applied), and the amount of models used for bagging ($m_{bagging}$). For a short explanation of bagging and regularization see Appendix A. Optimizing the input parameters and modeling settings can be done in two different ways. The first approach is to optimize the modeling performance by hand, changing inputs and modelings settings up until the point that the model performance (R^2_test) does not improve anymore. The second approach would be to develop an algorithm that automatically alters the inputs and modeling settings in such a way that the model performance increases the most.

For this research, it was chosen to apply both strategies. First, the linear regression models are improved by hand to get a feeling for which alterations of the inputs or modeling settings increase performance the most. The choosing of the best input parameters in the by-hand analysis is based on the three different strategies described in subsection 5.3.1. This by hand analysis is useful to get a first indication of the modeling performances that can be reached. Which in a later stadium can be used to validate the algorithm that automatically improves the model. When the algorithm significantly underperforms compared to the model improved by hand this indicates that the algorithm does not work properly and should be revised.

Also, the by hand analysis can identify if additions to the model like bagging and regularization add value to the model performance. As stated in subsection 2.2.1 regularization and bagging are likely to be of added value when the model overfits the training data. When it becomes clear in the by hand analysis that the models do not tend to overfit the training data and therefore do not add value these components (regularization and bagging) can be turned off in the optimization algorithm to reduce the required simulation time. In the next subsection, the developed algorithm to automatically improve the model performance is described. The performances and the resulting models of both the by hand and algorithm-based models are presented in chapter 6.

5.5.1. Improve the model automatically by algorithm

In this section, the algorithm that is developed, to automatically optimize the multivariate linear regression model, is described. As already mentioned in section 5.5 two things need to be optimized, namely the chosen model inputs and the bagging and regularization settings. The developed algorithm optimizes these in an iterative process. The algorithm starts with a certain beginning situation, which needs to be inputted, the inputs required to define the start situation are the chosen input parameters ($start\_input\_pars$) and the model settings ($start\_model\_settings$). The models' settings include the number of splits for k-fold splitting (n_kfold), two booleans to turn on and of both regularization and bagging (reg, bag), the regularization terms ($L1_{wt}$ and λ) and the amount of models used for bagging ($m_{bagging}$).

The chosen inputs ($input\_pars$) and model settings ($model\_settings$) are the two inputs that are required to train a corresponding ML model. In the algorithm, the chosen inputs and the model settings are updated every iteration in such a way that the model performance ($R2\_test$) is increased. The only exception is the number of splits for k-fold splitting, this model input is kept constant during the simulation. The iterative process repeats itself until there are no alterations left that further improve the model performance. When this point is reached the algorithm stops and the modeling settings ($model\_settings\_max$) and input parameters ($input\_pars\_max$) that are used in the last iteration are outputted together with the corresponding performance ($R2\_test\_max$).

For choosing the best set of input parameters the developed algorithm applies a strategy that is very similar to the stepwise linear regression method described in subsection 5.3.1. The only difference is that it is possible to start the analysis with any (combination) of input parameter(s) that is chosen by the user, instead of starting with only the most promising input parameter. However, when the algorithm is applied in this research only the most promising input parameters is chosen as the start input ($start\_input\_pars$). Another difference between standard stepwise regression and the developed algorithm is that the algorithm also includes the optimization of the modeling settings in the iterative process.

A flowchart of the algorithm is shown in Figure 5.2. The dotted lines in the flowchart represent data flows that only occur once, either in the first iteration or the last. All the solid lines represent the iteration loop that keeps iterating until the model performance stops increasing and the model results are outputted in the bottom right corner of the flowchart. Furthermore, parameters representing lists (multiple values) are underlined in the flowchart. To for example make a clear distinction between all the input parameters ($input\_pars$) and a single input parameter ($input\_par$) considered for removing or adding.

As can be seen in the top part of the flowchart the algorithm requires 6 inputs. Besides the two inputs already described above ($start\_input\_pars$ and $start\_model\_settings$) four other inputs are required. The first other input, is a threshold value for selecting potential model inputs ($threshold\_cor$). This value represents the minimum correlation value a parameter from the total input table needs to have with the dependent variable (y) to be considered in the algorithm. Choosing zero for this value means that all inputs are considered, whereas choosing one would mean not a single input is considered. This threshold is included to improve the modeling speed by excluding some weakly correlated input parameters. The logic for choosing the potential inputs (pot_inputs) for the algorithm only happens once in the beginning and is depicted in the bottom left corner of the flowchart.

After all potential inputs are identified the iteration loop can start. In the first block of the iteration loop, all potential inputs are one by one considered as new input for the model. The new potential input with the highest score ($R2\_test\_new$) is then compared to the current best known model score ($R2\_test$). If adding the new input resulted in a higher score the new input is added to the input parameter list and removed from the potential inputs list. The second block of the iteration loop does something very similar but instead of adding all potential inputs all input parameters that are currently used are removed one by one. If removing one of the inputs increases the model score this input is removed from the model and is no longer part of the simulation. The blocks so far described are together responsible for the first optimization goal (choosing the best input parameters).

The second part of the algorithm is responsible for finding the best model settings. First, the modeling settings for regularization are considered, by initially checking if the model performance increases in regulation is turned off. Afterward, no matter if regulation is on or off, it is checked whether changing the regularization parameters improves the model performance. This is where the fourth and fifth inputs come to play, the considered regularization types ($considered\_L1\_wt$) and the step size of increasing the penalty term lambda ($step\_size\_lambda$). For all regularization types that are considered the penalty term lambda is increased until the model performance does not increase anymore. When all the different regularization types are done the score of the best performing regularization model (max_R2_reg) is compared to the current best score ($R2\_test$). If the new regularization model outperforms the current model the regularization terms are updated.

After the regularization, in the final part of the algorithm, bagging is considered in a very similar way. For this part of the algorithm the sixth final input is required, the step size of amount of models used for bagging. As for the lambda, the amount of models used for bagging is increased until the model performance stops increasing. The highest model performance again is compared to the current best score and when the bagging model is better the model input n_models is updated.

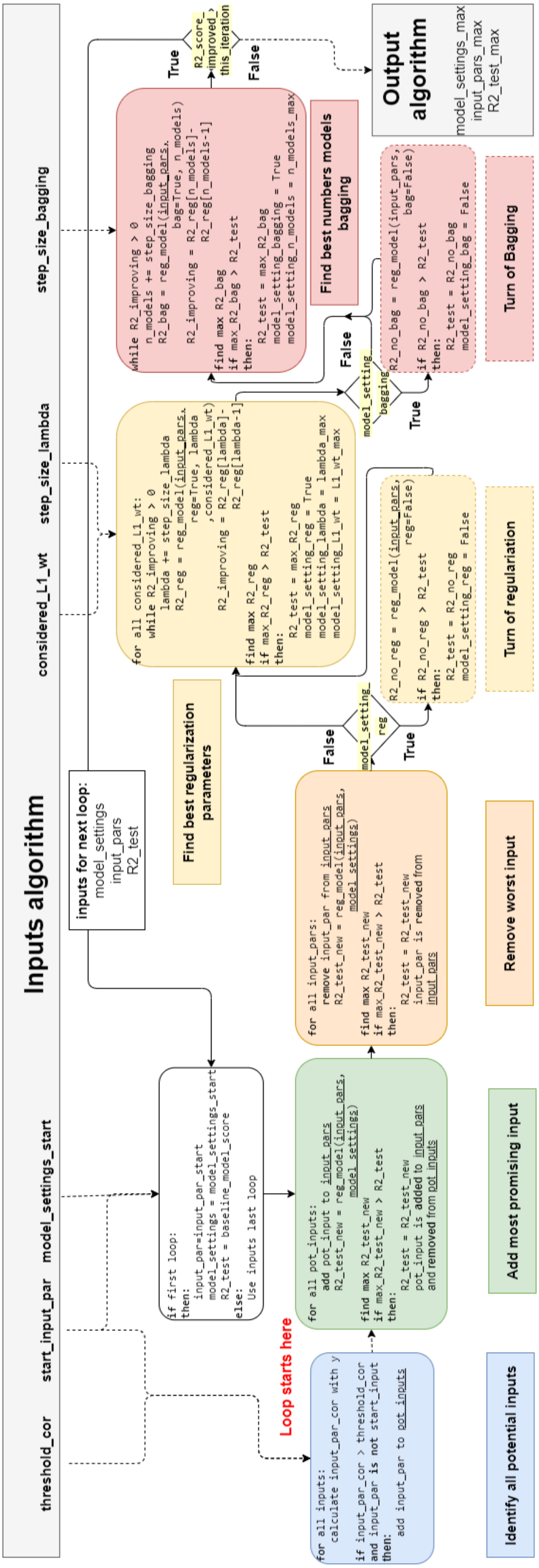

Inputs algorithm
threshold_cor
start_input_par
model_settings_start
considered_L1_wt
step_size_lambda
step_size_bagging
Loop starts here
for all inputs:
calculate input_par_cor with y
if input_par_cor > threshold_cor
and input_par is not start_input
then:
add input_par to pot_inputs
Identify all potential inputs
if first loop:
then:
input_par=input_par_start
model_settings = model_settings_start
R2_test = baseline_model_score
else:
Use inputs last loop
for all pot_inputs:
add pot_input to input_pars
R2_test_new = reg_model(input_pars, model_settings)
find max R2_test_new
if max_R2_test_new > R2_test
then:
R2_test = R2_test_new
pot_input is added to input_pars and removed from pot_inputs
Add most promising input
for all input_pars:
remove input_par from input_pars
R2_test_new = reg_model(input_pars, model_settings)
find max R2_test_new
if max_R2_test_new > R2_test
then:
R2_test = R2_test_new
input_par is removed from input_pars
Remove worst input
inputs for next loop:
model_settings
input_pars
R2_test
model_setting_reg
True
False
R2_no_reg = reg_model(input_pars, reg=False)
if R2_no_reg > R2_test
then:
R2_test = R2_no_reg
model_setting_reg = False
Turn of regulariation
for all considered_L1_wt:
while R2_improving > 0
lambda += step_size_lambda
R2_reg = reg_model(input_pars, reg=True, lambda ,considered_L1_wt)
R2_improving = R2_reg[lambda]- R2_reg[lambda-1]
find max R2_reg
if max R2_reg > R2_test
then:
R2_test = max_R2_reg
model_setting_reg = True
model_setting_lambda = lambda_max
model_setting_L1_wt = L1_wt_max
Find best regularization parameters
model_setting_bagging
True
False
R2_no_bag = reg_model(input_pars, bag=False)
if R2_no_bag > R2_test
then:
R2_test = R2_no_bag
model_setting_bag = False
Turn of Bagging
while R2_improving > 0
n_models += step_size_bagging
R2_bag = reg_model(input_pars, bag=True, n_models)
R2_improving = R2_reg[n_models]- R2_reg[n_models-1]
find max R2_bag
if max_R2_bag > R2_test
then:
R2_test = max_R2_bag
model_setting_bagging = True
model_setting_n_models = n_models_max
Find best numbers models bagging
R2_score_improved_this_iteration
True
False
Output algorithm
model_settings_max
input_pars_max
R2_test_max

Figure 5.2: Flowchart of algorithm to automatically optimize model performance

III

Part 3 Research outcome

6 The Resulting Models and their Performance

In this chapter, the optimization process and the final results of the five different linear regression models are described. As mentioned in section 5.5, the models are first optimized by hand, before an optimizing algorithm, see subsection 5.5.1 for algorithm explanation, is used. First, the two "normal" model using only gas and water data are described in section 6.1 and section 6.2 respectively. Then in section 6.3 the models using both water and gas data to make predictions are presented. Afterward, in section 6.4 the results of the dummy analysis are discussed. Finnaly, the formulas and some characteristics of the best models for predicting water and gas are presented in section 6.5. Finally, the consequences of all these results for DH cost modeling are discussed in section 6.6.

6.1. The gas model

In this subsection, the modeling results for the gas model are described. The subsection is subdivided into three parts, first in subsection 6.1.1 the results of the gas models that were tuned by hand are discussed. Second, subsection 6.1.2 describes the results of the algorithm to automatically improve the gas model performance. Finally, non-linear transformation(s) of the input parameters and the corresponding modeling performances are presented in subsection 6.1.3.

6.1.1. Choosing input parameters and model settings by hand

When optimizing the gas model by hand it was chosen to first focus on choosing the right input parameters before considering bagging and regularization. As mentioned in subsection 5.3.1 there are three different approaches for choosing input parameters. The first method, forward regression, was applied first. Normally when applying this strategy one would stop adding input parameters as soon as the performance stops increasing. However since a goal of the by hand analysis is to get a feeling of the modeling process, it is chosen to keep adding input parameters until the model performance reduces significantly.

The question that remains is how to select the most promising input parameter. As shown in Table 5.6 there are two ways to identify high potential parameters, the correlation of a certain input (x) with the dependent variable (y) and the Tolerance value of a potential input, see subsection A.0.3. The most promising input parameter was selected by simply multiplying the correlation and the Tolerance value of all potential input parameters and selecting the parameter with the highest multiplication result. The input parameter that was included first is diameter since it represents the baseline model.

Furthermore, it was chosen to set n_kfold to 5 which means that the test set contains either 10 or 11 projects. The results from this first optimization approach are presented in Table 6.1. In the first row, the inputs that are added to the model are presented with their correlation times tolerance value between brackets. In the second row, the average training score (goodness-of-fit) is given. Since the training score is more stable and not used for optimizing directly, only the average of the n trained models is included in the table. For the test score (**prediction accuracy**) the minimum, maximum, and average scores are included in row 3,4 and 5. As stated in section 5.2 the average prediction accuracy of a model is used as the main scoring criteria.

The minimum and maximum values of R^2 are added to the table to give an indication of the spread of the different R^2 scores.

The first serious reduction in model performance can be seen when the input parameter "Tunnel" is added. That is why the input Tunnel is removed again before the next input ("Distance to road") is included. Finally, after the 8th input is added ("Percentage trees") it is concluded that the model performance has reduced that much compared to the third input ("Soil replacement") that adding more inputs is not sensible anymore.

Table 6.1: Model scores (R^2) gas model, including inputs selected based on correlation and tolerance value selection criteria between brackets

Model settings: n_kfold=5, Regularization=False, Bagging=False

Input added:	**Diameter**	**Footprint building (0.45)**	**Soil replace-ment (0.32)**	Tunnel (0.32) removed after	**Distance to road (0.24)**	**Total length (0.21)**	**Pollution category (0.19)**	**Woning overig (0.18)**	**Percentage trees (0.17)**
train avg	0.19	0.37	0.4	0.44	0.41	0.44	0.45	0.45	0.45
test min	-0.04	-0.1	0.05	-0.55	-0.3	-0.28	-0.27	-0.3	-0.36
test max	0.35	0.53	0.5	0.43	0.53	0.59	0.5	0.49	0.48
Test AVG	**0.16**	**0.15**	**0.22**	**0.09**	**0.14**	**0.12**	**0.11**	**0.09**	**0.01**

After applying the first method to select the best input parameters the second method, backward regression, was applied. In the second method, the model is first trained using all the inputs, and the worst-performing input parameter is removed until the model performance stops increasing. However, since the total amount of potential inputs is very high it is chosen to only consider the input parameters which have an (absolute) correlation with the dependent variable which is higher than 0.2. All the inputs that meet this required are shown in Table 6.2.

Table 6.2: Correlations potential input parameters for the Gas model with construction costs

Input parameter	Correlation
Footprint building[%]	0.45
AVG diameter	0.43
Inhabitants	0.41
Soil replacement	0.38
Tunnel	0.37
Distance to road	0.35
Pollution category	0.32
Percentage trees	0.29
Woning overig	0.26
Network operator 1	0.25
Network operator 2	0.24
Total length	0.24
PVC_medium	0.20

As for the first method is it chosen to keep removing inputs even though the performance is decreasing to check if the performance might increase again after a local maximum. The inputs are removed based on their t-statistic, the input parameter with the highest t-statistic is removed.

To speed up the process in the first two iterations two inputs are removed at once since all these parameters have relatively high t values. The results of the second optimization strategy are shown in Table 6.3. The table is very similar to the table showing the results of the first method. The main difference is that in the first row now the removed input parameter(s) with the corresponding t-statistic(s) are shown. As can be seen in the table the model performs best when all inputs except: "Total_length", "Pollution_category", "Inhabitants", "AVG_diameter" and "Network_opr_1" are removed. It can be seen that removing based on t-stat results in better model performances than adding based on tolerance value and correlation.

Table 6.3: Model scores (R^2) gas model, removing inputs based on t-statistic value

Model settings: n_kfold=5, Regularization=False, Bagging=False

Input Removed:	**all included**	**Network opr 2 (0.85), Woning overig (0.80)**	**Soil replacement (0.79), Distance to road (0.75)**	**Percentage trees (0.66)**	**PVC_ medium (0.34)**	**Footprint building (0.33)**	**Total length (0.06)**	**Pollution category (0.03)**
train avg	0.63	0.62	0.62	0.61	0.6	0.59	0.54	0.47
test min	-0.39	-0.34	-0.29	-0.28	-0.05	-0.1	0.11	-0.13
test max	0.56	0.54	0.58	0.58	0.67	0.69	0.53	0.56
Test AVG	**0.18**	**0.2**	**0.22**	**0.28**	**0.36**	**0.37**	**0.34**	**0.25**

After trying two methods for choosing the best input parameters the effect of applying regularization is studied. For this analysis the four best performing sets of input parameters so far identified are used as inputs. The model setting "regularization" is turned on and three different values for lambda are inputted. This analysis is done for L1 regularization, L2 regularization, and combined regularization, using $L1\_wt = 0.5$. The results of the three analyses were very similar, but the combined regularization resulted in the highest overall performance score (0.38) and is therefor shown below in Table 6.4. The tables with the results from the other analysis are given in Appendix C. The results in the table indicate that using L2/L1 regularization has the potential to slightly improve the model performance of the gas model.

Table 6.4: Model scores (R^2) gas model, trying different lambda's for L2/L1 regularization

Model settings: n_kfold=5, Regularization=True, reg_L1_wt = 0.5, Bagging=False

Inputs Model:	Total length, pollution, inhabitants, diameter, network opr 1			Total length, pollution, inhabitants, diameter, network opr1, footprint building			Total length, pollution, inhabitants, diameter, network opr 1, footprint building, PVC_medium			Total length, pollution, inhabitants, diameter, network opr 1, footprint building, PVC_medium, Percentage trees		
Lambda	**0.1**	**0.25**	**0.5**	**0.1**	**0.25**	**0.5**	**0.1**	**0.25**	**0.5**	**0.1**	**0.25**	**0.5**
train avg	0.57	0.55	0.52	0.57	0.55	0.52	0.59	0.56	0.53	0.59	0.57	0.54
test min	-0.06	-0.07	-0.11	-0.06	-0.07	-0.11	-0.31	-0.31	-0.32	-0.32	-0.31	-0.33
test max	0.65	0.61	0.60	0.65	0.61	0.60	0.62	0.61	0.61	0.57	0.57	0.57
Test AVG	**0.37**	**0.36**	**0.32**	**0.38**	**0.36**	**0.32**	**0.29**	**0.28**	**0.25**	**0.25**	**0.24**	**0.22**

Besides analyzing the effects of regularization the effects of bagging are also examined. The same four input sets as for regularisation are used and again three different values for the tuning factor (number of models) are inputted. The results of the bagging analysis can be seen in Table 6.5. From the table, it can be concluded that also bagging has the potential to slightly increase the model performance of the Gas model.

Table 6.5: Model scores (R^2) gas model, trying different number of models bagging

Model settings: n_kfold=5, Regularization=False, Bagging=True

Inputs Model:	Total length, pollutions inhabitants, diameter, network opr 1			Total length, pollution, inhabitants, diameter, network opr1, footprint building			Total length, pollution, inhabitants, diameter, network opr 1, footprint building, PVC_medium			Total length, pollution, inhabitants, diameter, network opr 1, footprint building, PVC_medium, Percentage trees		
N_models	**3**	**8**	**12**	**3**	**8**	**12**	**3**	**8**	**12**	**3**	**8**	**12**
train avg	0.56	0.57	0.58	0.56	0.58	0.59	0.56	0.59	0.60	0.56	0.60	0.61
test min	-0.04	-0.08	-0.14	-0.07	-0.06	-0.12	-0.19	-0.39	-0.63	-0.20	-0.49	-0.70
test max	0.69	0.67	0.68	0.69	0.67	0.66	0.56	0.62	0.64	0.49	0.62	0.64
Test AVG	**0.39**	**0.38**	**0.36**	**0.38**	**0.38**	**0.35**	**0.27**	**0.26**	**0.21**	**0.22**	**0.21**	**0.15**

The final analysis that is conducted, before the gas data is used in the algorithm to automatically improve the model performance, is the effect of combining bagging and regularization on the model performance. From the four sets of input parameters used so far only the first two, which have shown the most promising results, are used in this final analysis. Also for regularization, the most promising settings from the separate regularization analysis are used. Meaning that $L1\_wt$ is set to 0.5 and for $lambda$ 0.1 and 0.25 are inputted. The results of this combined analysis are shown in Table 6.6. As can be seen in the table combining regularization and bagging does not further improve the modeling performance of the gas model.

Table 6.6: Model scores (R^2) gas model, combining bagging and regularization

Model settings: n_kfold=5, Regularization=True, Bagging=True

Inputs Model:	Total length, pollution, inhabitants, diameter, network opr 1			Total length, pollution, inhabitants, diameter, network opr1, footprint building			Total length, pollution, inhabitants, diameter, network opr 1			Total length, pollution, inhabitants, diameter, network opr1, footprint building		
N_models	**3**	**8**	**12**	**3**	**8**	**12**	**3**	**8**	**12**	**3**	**8**	**12**
Lambda	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**
train avg	0.56	0.57	0.58	0.56	0.58	0.59	0.56	0.59	0.60	0.56	0.60	0.61
test min	-0.04	-0.08	-0.14	-0.07	-0.06	-0.12	-0.19	-0.39	-0.63	-0.20	-0.49	-0.70
test max	0.69	0.67	0.68	0.69	0.67	0.66	0.56	0.62	0.64	0.49	0.62	0.64
Test AVG	**0.39**	**0.38**	**0.36**	**0.38**	**0.38**	**0.35**	**0.27**	**0.26**	**0.21**	**0.22**	**0.21**	**0.15**

6.1.2. Improving the model with the algorithm

In this part of the analysis, the algorithm described in subsection 5.5.1 is used to automatically find the best combination of input parameters and modeling settings. Based on the results of the by hand analysis it is chosen to include all blocks of the algorithm, and thus include bagging and regularization in the optimization process. As can be seen in the flowchart in subsection 5.5.1 the algorithm requires 5 inputs. The first input is the correlation threshold used for identifying potential input parameters. This threshold is set to 0.1. By using 0.1 more inputs are considered than in the by hand analysis but still almost half the inputs are not considered to decrease the modeling time. The second algorithm input is, the input parameters used in the first iteration. It is chosen to start the analysis by only using the $AVG\_diameter$. As the third input, the initial modeling settings are required. In the first iteration regularization and bagging are turned off. The number of splits in k-fold splitting, which does not change over the iterations, is again set to 5 to create a fair comparison between the by hand modeling results and the algorithm. Finally, the same three regularization types (L1, L2, L1/L2) are considered and the step size for lambda and the number of models for bagging respectively are 0.05 and 2. The results of the algorithm are shown in Table 6.7, as can be seen, the algorithm was able to increase the model performance from 0.39 up to 0.45 indicating that the algorithm performs properly.

Table 6.7: Result algorithm gas model

Gas model

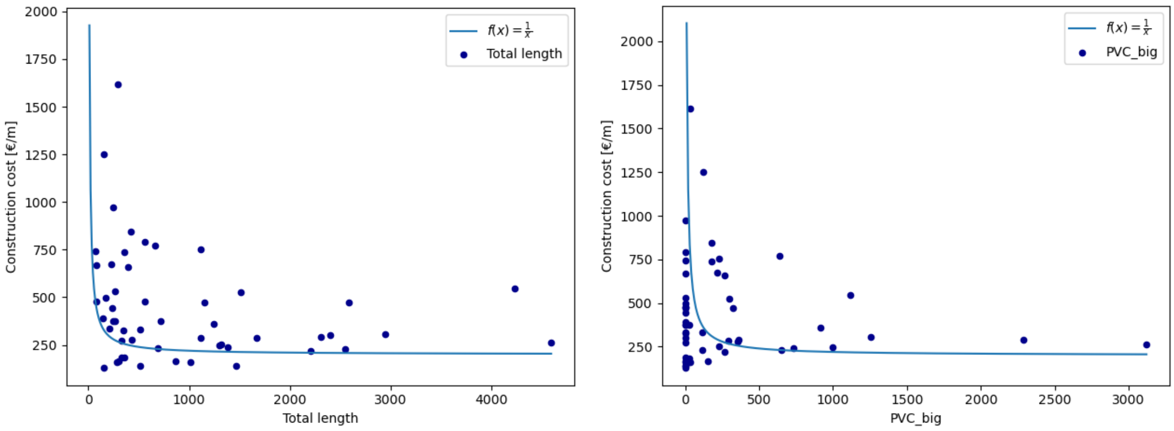
Algorithm inputs		**Algorithm outputs**	
start_input_par	[AVG diameter]	**input_par**	[AVG diameter, Pollution category, Woning overig, Network opr 1, PVC_big, PE_medium, AVG maxspeed]
n_kfold	5		
Reg_start	False		
Bagging_start	False	**Reg**	True
Threshold_cor	0.1	**Lambda**	0.1
Considered_L1_wt	[0, 0.5, 1]	**L1_wt**	1
Lambda_step	0.05	**Bagging**	False
n_models_bagging_step	2	**n_models_bagging**	[-]
R2 test AVG: 0.45			

6.1.3. Non linear transformations of input parameters

When only choosing the best inputs from the "normal" input table the assumption is made that the relations between the inputs and the dependent variable (y) are all linear. In practice, however, it is unlikely that all the relations between the inputs and the dependent variable are in fact linear. For this reason, it is worthwhile to check which inputs seem to have non-linear relations with the dependent variable. When a specific non-linear relation type can be identified the original input can be transformed and the model performance is likely to increase. Because there are a lot of potential input parameters and it is quite time consuming to check all of them for potential non-linear relations it is chosen to only study a selection of the inputs.

The inputs that are chosen to study are all the inputs from the best performing model found using the algorithm and also the inputs from the best performing model found in the by hand analysis. Besides the inputs from the best performing models also some extra inputs are checked because these inputs were considered promising and yet did not show up in the best performing models. Only plots for numerical input parameters are studied since for categorical values it is almost impossible to find a non-linear patern since they have very limited spread on the x-axis. Combining the above-described inputs resulted in the following list of input parameters that were studied for non-linear relations with the construction cost per meter for gas projects: *Total length, Inhabitants, AVG diameter, PVC_big, PE_medium, AVG maxspeed, Footprint building, Distance to road, Percentage trees, Building year diff med and mean, Years ago and Percentage asphalt.*

For all these input parameters scatter plots were generated that were analyzed by hand to find non-linear patterns. Out of all these input parameters two were found that showed a clear non-linear relation. The scatter plots of these two input parameters (Total length, PVC_big) are shown in Figure 6.1. The rest of all the scatter plots can be found in Appendix C.



(a) Total length scatter plot

(b) PVC_big scatter plot

Figure 6.1: Scatter plots non-linear input parameters gas

The graphs in Figure 6.1 show that, instead of a linear relation the Total length and PVC_big seem to have a $\frac{1}{x}$ relation with the dependent variable (y). Because of this $\frac{1}{Total\_length}$ and $\frac{1}{PVC\_big}$ are added to the total input table. Using this new total input table the algorithm is used again to find the best performing model, to check if the algorithm will include the transformed inputs and if so, if these inputs increase the model performance. The results of this analysis are shown in Table 6.8, as can be seen, the non-linear input $\frac{1}{Total\_length}$ is included in the final model and the model performance increased by 0.01 to 0.46.

Table 6.8: Results algorithm gas model with non linear transformations

Gas model			
Algorithm inputs		**Algorithm outputs**	
start_input_par	[AVG diameter]	**input_par**	['AVG diameter', 'Pollution category', 'Woning overig', 'Waterutility 2', 'PVC_big', 'PE_medium', '1/Total length']
n_kfold	5		
Reg_start	False		
Bagging_start	False	**Reg**	True
Threshold_cor	0.1	**Lambda**	0.1
Considered_L1_wt	[0, 0.5, 1]	**L1_wt**	1
Lambda_step	0.05	**Bagging**	False
n_models_bagging_step	5[1]	**n_models_bagging**	[-]
R2 test AVG: 0.46			

6.2. The water model

In this subsection, the modeling results for the water model are presented. This section is divided into the same three subsections as the gas model section. First, the model is optimized by hand, then the optimization algorithm is used and finally, non-linear transformations on input parameters are performed.

6.2.1. Choosing input parameters and model settings by hand

For the by hand optimizing process of the water model, a similar approach is applied as for the gas model described above. First, the "best" input parameters are studied in two different ways, and then the effect of regularization and bagging is analyzed. As for the gas model, first, model inputs are added to the model based on their tolerance and correlation value, starting with the diameter (forward regression). A difference compared to the gas model is that for the water model n_kfold is set to 6 because the total amount of datapoints is higher.

The model performances resulting from this approach are shown in Table 6.9. As can be seen in the table, the model performances of the water model are significantly lower than the model performances of the gas model. This is interesting because based on the number of projects it would be expected that the water model has a higher model performance. There could be multiple reasons for this, some of which are discussed in chapter 7. In this chapter, the focus lies on improving the performances instead of explaining the differences in model performances between different models.

Table 6.9: Model scores (R^2) water model, including inputs based on correlation and tolerance value

Model settings: n_kfold=6, Regularization=False, Bagging=False

Input added:	**Diameter**	**Total length (0.38)**	**Footprint Building (0.32)**	**Road overig (0.30)**	**Water utility 1 (0.23)**	**Dewatering (0.23)**	**Pollution (0.2)**	**Water utility 2 (0.19)**	**Footprint road (0.16)**
train avg	0.34	0.45	0.52	0.54	0.62	0.62	0.65	0.66	0.67
test min	-0.78	-0.72	-1.18	-0.75	-1.7	-1.67	-1.39	-1.29	-1.29
test max	0.36	0.41	0.5	0.53	0.54	0.55	0.53	0.51	0.44
Test AVG	**-0.17**	**0.02**	**-0.03**	**0**	**0**	**0.03**	**0.04**	**0.06**	**0.01**

For the second approach (backward regression), for choosing the best inputs, again only the input parameters that have a correlation of at least 0.2 are considered. The potential inputs that have a correlation of at least 0.2 with the price per meter of water networks are shown in Table 6.10.

Table 6.10: Correlations potential input parameters for the Water model

Input parameter	**Correlation**
AVG diameter	0.59
Total length	0.39
Road Overig	0.38
Footprint building[%]	0.34
Water utility 1	0.31
Water utility 2	0.31
Dewatering	0.29
PVC_medium	0.24
Footprint road [%]	0.23
Pollution category	0.22
Distance to road	0.21
PVC_small	0.21
Planned time	0.21

Starting with all the input parameters from Table 6.10, inputs are removed based on their t-statistic. In some iterations, two parameters are removed at the same time to speed up the process. Also in iteration five two potential input parameters have the same t-statistic and are therefore removed in the same iteration. The

[1] Model setting number of models bagging changed because of better results in water analysis, see subsection 6.2.2

results from this analysis are shown in Table 6.11. In contrast to the gas model for the water model the best performance for both input parameter selecting strategies are the same (0.06).

Table 6.11: Model scores (R^2) water model, removing inputs based on t-statistic value

Model settings: n_kfold=6, Regularization=False, Bagging=False

Input removed:	**all parameters included**	**Distance road (0.80), total length (0.73)**	**PVC_medium (0.69), Footprint road (0.56)**	**PVC_small (0.53)**	**Planned time (0.40)**	**Dewatering (0.39)**	**Water utility 2 (0.31)**	**Pollution category (0.18)**
train avg	0.68	0.68	0.67	0.67	0.66	0.65	0.64	0.61
test min	-2.7	-2.4	-0.99	-0.96	-1.12	-1.18	-1.2	-1.46
test max	0.43	0.43	0.52	0.52	0.5	0.49	0.49	0.55
Test AVG	**-0.37**	**-0.29**	**0.04**	**0.06**	**0.07**	**0.05**	**0.02**	**0**

To analyze the effects of regularization and bagging again the four best performing sets of input parameters are used. In contrast to the gas model, both bagging and regularization do increase the modeling performance of the water model significantly. The modeling approach which increases the modeling performance the most is L2 regularization. The results from this analysis are shown in Table 6.12. The results from the other two regularization analysis and the bagging analysis can be found in Appendix D.

Table 6.12: Model scores (R^2) water model, trying different lambda's for L2 regularization

Model settings: n_kfold=6, Regularization=True, reg_L1_wt = 0, Bagging=False

Inputs model:	Diameter, footprint building, Water utility 1, road overig, Pollution category, Water utility 2, Dewatering			Diameter, footprint building, Water utility 1, road overig, Pollution category, Water utility 2, Dewatering, Planned time			Diameter, Total length, Footprint building, Road overig, Dewatering, Water utility 1, Pollutioin category, Water utility 2			Diameter, footprint building, Water utility 1, Road overig, Pollution category, Water utility 2		
Lambda	**0.1**	**0.25**	**0.5**	**0.1**	**0.25**	**0.5**	**0.1**	**0.25**	**0.5**	**0.1**	**0.25**	**0.5**
train avg	0.58	0.53	0.48	0.57	0.53	0.50	0.59	0.55	0.51	0.56	0.52	0.47
test min	-0.03	-0.10	-0.20	-0.40	-0.50	-0.61	-0.29	-0.23	-0.31	-0.02	-0.11	-0.21
test max	0.51	0.53	0.55	0.51	0.52	0.52	0.52	0.56	0.59	0.49	0.52	0.54
Test AVG	**0.23**	**0.21**	**0.14**	**0.13**	**0.10**	**0.08**	**0.23**	**0.20**	**0.15**	**0.23**	**0.20**	**0.13**

Table 6.13: Model scores (R^2) water model, combining bagging and regularization

Model settings: n_kfold=6, Regularization=True, reg_L1_wt = 0, Bagging=True

Inputs model:	Diameter, footprint building, Water utility 1, road overig, Pollution category, Water utility 2, Dewatering				Diameter, Total length, Footprint building, Road overig, Dewatering, Water utility 1, Pollutioin category, Water utility 2				Diameter, footprint building, Water utility 1, Road overig, Pollution category, Water utility 2			
Lambda	**0.1**	**0.1**	**0.25**	**0.25**	**0.1**	**0.1**	**0.25**	**0.25**	**0.1**	**0.1**	**0.25**	**0.25**
n_models	**20**	**40**	**20**	**40**	**20**	**40**	**20**	**40**	**20**	**40**	**20**	**40**
train avg	0.57	0.57	0.52	0.52	0.58	0.58	0.54	0.54	0.55	0.55	0.51	0.50
test min	-0.04	0.02	-0.15	-0.11	-0.18	-0.16	-0.18	-0.14	-0.06	0.03	-0.14	-0.10
test max	0.49	0.51	0.52	0.54	0.51	0.53	0.54	0.56	0.48	0.52	0.50	0.53
Test AVG	**0.22**	**0.26**	**0.19**	**0.22**	**0.22**	**0.26**	**0.18**	**0.22**	**0.23**	**0.27**	**0.19**	**0.22**

The final analysis that again is performed by hand is combining the most promising regularization settings with the most promising bagging settings to check whether this further improves the modeling performance. As already stated above the most promising regularization for the water model was L2 regularization. In Table 6.12 it can be seen that the best performing models are the models in the first, the third, and the fourth input set with $lambda$ equal to 0.1 or 0.25. That is why these models are combined with bagging, for bagging the most promising number of bagging models were 20 and 40. The results of the combined analysis can be seen in Table 6.13. As can be concluded from the table, that combining bagging and regularization does improve the model performance for the water model from 0.23 (only L2 regularization) up to 0.27 (bagging and L2 regularization).

6.2.2. Improving the model with the algorithm

In this subsection, the results of the optimization algorithm applied to the water model are discussed. Based on the by hand analysis, it is concluded that both regularization and bagging have to potential to significantly improve the model performance of the water model. Therefore, again, it was chosen to include both regularization and bagging in the algorithm. The initial inputs of the water model are the same as for the gas model except for the amount of splits used for k-fold splitting. As for the by hand analysis of the water model, the amount of splits is set to 6 instead of 5 because the dataset is slightly bigger. The results of the optimization algorithm for the water model are presented in Table 6.14.

Table 6.14: Results algorithm water model

Water model			
Algorithm inputs		**Algorithm outputs**	
start_input_par	[AVG diameter]	**input_par**	[AVG diameter, Road width, Total length, PVC_big, Water utility 2, Pollution category]
n_kfold	6		
Reg_start	False		
Bagging_start	False	**Reg**	True
Threshold_cor	0.1	**Lambda**	0.2
Considered_L1_wt	[0, 0.5, 1]	**L1_wt**	1
Lambda_step	0.05	**Bagging**	False
n_models_bagging_step	2	**n_models_bagging**	[-]
R2 test AVG: 0.24			

In contrast to the gas model performance resulting from the algorithm, the algorithm does not improve the model performance of the water model. The model performance found by the algorithm (0.24) is 0.03 lower than the best model found by hand. This is likely to be caused by how the number of models for bagging is optimized. Starting from 0 the number of models is increased by 2 until the model performance does not increase anymore. The best model found so far uses forty models for bagging, it is quite likely that the optimizing algorithm never reaches such high numbers since the performance decreases ones in the process.

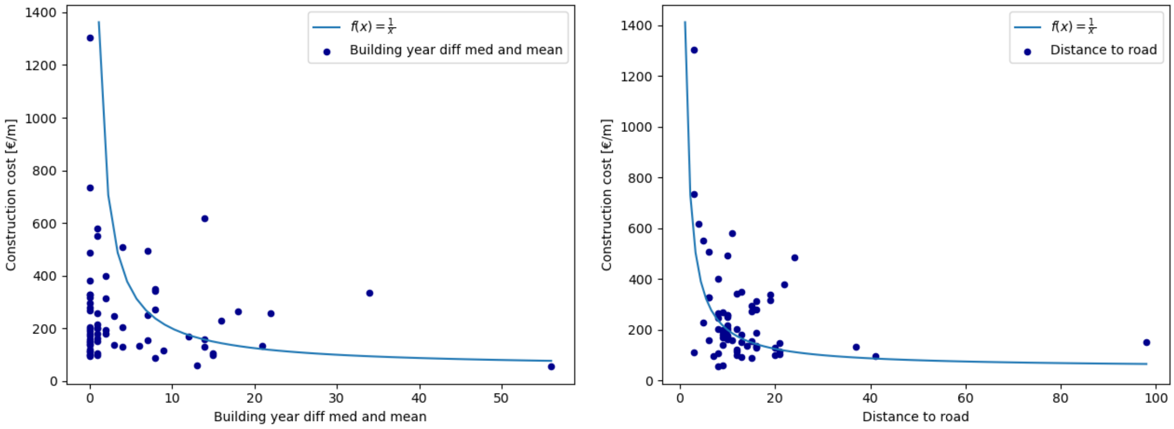
Table 6.15: Results algorithm water model bigger step size bagging models

Water model			
Algorithm inputs		**Algorithm outputs**	
start_input_par	[AVG diameter]	**input_par**	[’Water utility 1’, ’AVG diameter’, ’Water utility 2’, ’Pollution category’, ’PVC_big’, ’open verharding [%]’]
n_kfold	6		
Reg_start	False		
Bagging_start	False	**Reg**	True
Threshold_cor	0.1	**Lambda**	0.1
Considered_L1_wt	[0, 0.5, 1]	**L1_wt**	0
Lambda_step	0.05	**Bagging**	True
n_models_bagging_step	5	**n_models_bagging**	30
R2 test AVG: 0.28			

Because of this reason, it is chosen to also run a simulation with a higher step size for the number of models used for bagging to see if this improves the algorithm performance. The results of this analysis are shown in Table 6.15. It can be seen that indeed the performance of the algorithm increased when the step size was set to 5 and therefore this step size is used from now on. Even though the improved algorithm did not significantly increase (only 0.1) the model performance compared to the already found best-performing model shown in Table 6.13. The algorithm seems to work when the step size is set to five.

6.2.3. Non linear transformations of input parameters

As for the gas model, also non-linear transformations are considered for the water model. To check whether these relations exist and if performing non-linear transformation to input parameters to compensate for these non-linear relations will improve the model performance. Again not all the input parameters are checked by hand to identify potential transformations but only promising input parameters are considered. Parameters are considered promising either because they are used as input for the best performing models (algorithm or by hand), or because they are not used as inputs yet when it was expected that these inputs could be of significant added value.



(a) Building year diff med and mean scatter plot

(b) Distance to road scatter plot

Figure 6.2: Scatter plots non-linear input parameters water

The two inputs that are considered but are not included in the models the best performing models are: *Building year diff med and mean and Distance to road.* The inputs that are considered, again only numerical parameters, because they are used in the best performing models are: *AVG diameter, Footprint building, Total length, PVC_big and Road width.* When looking at the scatter plots of all these inputs parameters four potential non-linear relations were found. Total length and PVC_big showed the same $\frac{1}{X}$ relation as was found for the gas model. But also the building year diff med and mean and Distance to road show potential for a $\frac{1}{x}$ relation. The scatterplots for these last two input parameters are shown in Figure 6.2. The scatter plots of all the other considered input parameters again are available in Appendix D.

Table 6.16: Results algorithm water model with non linear transformations

Water model			
Algorithm inputs		**Algorithm outputs**	
start_input_par	[AVG diameter]	**input_par**	[’AVG diameter’, ’1/Total length’, ’Water utility 2’, ’Pollution category’ , ’PE_small’, ’Water utility 1’]
n_kfold	6		
Reg_start	False		
Bagging_start	False	**Reg**	False
Threshold_cor	0.1	**Lambda**	[-]
Considered_L1_wt	[0, 0.5, 1]	**L1_wt**	[-]
Lambda_step	0.1	**Bagging**	True
n_models_bagging_step	5	**n_models_bagging**	30
R2 test AVG: 0.43			

The four non-linear transformations of the input parameters are added to the total input table and then the algorithm is used again to find the best performing model. The results of this analysis can be seen in Table 6.16. As can be seen in the table the model performance increased significantly when $\frac{1}{Total\_length}$ is used as input instead of the normal $Total\_length$, now having an $R2\_test$ score of 0.43 which is 0.15 than the best water model found so far.

6.3. The combined models

After looking at the water and gas models separately in this subsection the water and gas projects are combined in one new bigger database. When adding the two databases together two extra columns are added, one column containing a 1 for every water project and one column containing a 1 for every gas project. These columns are added so that the difference in project type can be used as input for the ML model.

This combined database is used to train two different models, a model that is scored on prediction gas and a model that is scored on predicting water. This is done by altering the k-fold splitting process as described in subsection A.0.1. Instead of randomly splitting the entire database in n subsets, the database is first separated again into a database containing water projects and a database containing gas projects. When the combined model for gas is trained, the gas data is split into n batches, of which every batch is used ones for testing and $n-1$ times for training, like normal k-fold splitting. The difference compared to the normal gas model is that all the water projects are added to the training set after the splitting is done for all n training sets. By splitting the data like this, the size of the training set is increased while the test set remains the same. For the combined water model exactly the same trick is applied, only this time the water data is used for testing. The idea behind this is that model performance could increase because the water and gas projects are similar. Meaning that some information that is present in the water data could be of added value for the gas model and visa versa.

6.3.1. Combined gas model

To validate whether including water data in the training process of the gas model increases the modeling performance the training algorithm is used again with the same settings as for the normal gas model. Also, the same non-linear input parameter transformations as described in subsection 6.1.3 are applied to have a fair comparison. The only difference is the chosen starting input parameters, instead of only using the AVG diameter (baseline model), also the new column indicating which projects are water projects is included from the start. Adding this extra column from the beginning is done because the combined model does not function properly without this extra input parameter.

The characteristics of the resulting model are presented in Table 6.17, as can be seen, the performance of the gas model increased by 0.7 up to 0.53 when the water data is also used for training.

Table 6.17: Results algorithm combined gas model with non linear transformations

Combined Gas model			
Algorithm inputs		**Algorithm outputs**	
start_input_par	[AVG diameter, Water]	**input_par**	[’AVG diameter’, ’Water’, ’Network opr 1’, ’1/Total length’, ’Pollution category’, ’PVC_big’, ’Woning Vrijstaand’, ’PVC_medium’, ’PE_medium’, ’Years ago’, ’other_medium’, ’AVG maxspeed’]
n_kfold	5		
Reg_start	False		
Bagging_start	False	**Reg**	False
Threshold_cor	0.1	**Lambda**	[-]
Considered_L1_wt	[0, 0.5, 1]	**L1_wt**	[-]
Lambda_step	0.1	**Bagging**	True
n_models_bagging_step	5	**n_models_bagging**	15
R2 test AVG: 0.53			

6.3.2. Combined water model

For training the combined water model also the same non-linear input transformations and algorithm inputs as used for the normal water model are applied to have a fair comparison. Again only the starting input parameters are different because not just the AVG diameter but also the column indicating the gas projects is included. The results of this analysis are depicted in Table 6.18.

Unfortunately using the gas data as extra training data did not improve the modeling performance of the water model. Compared to the best performing water model found so far the combined model has a performance which is 0.7 lower. This could be because the water database is bigger than the gas database.

Adding the 68 water projects to the gas training data set which contains 42 or 41 data points might be more beneficial than adding the 52 gas projects to the water training data set which contains 56 or 57 data points.

Table 6.18: Results algorithm combined water model with non linear transformations

Combined Water model

Algorithm inputs		**Algorithm outputs**	
start_input_par	[AVG diameter, Gas]	**input_par**	[’AVG diameter’, ’Gas’, ’1/Total length’, ’Percentage asfalt’, ’Pollution category’, ’Water utility 2’, ’PE_small’]
n_kfold	6		
Reg_start	False		
Bagging_start	False	**Reg**	False
Threshold_cor	0.1	**Lambda**	[-]
Considered_L1_wt	[0, 0.5, 1]	**L1_wt**	[-]
Lambda_step	0.1	**Bagging**	True
n_models_bagging_step	5	**n_models_bagging**	10
R2 test AVG: 0.36			

6.4. The dummy model

As mentioned in section 5.4 dummy datasets are created in three different levels of complexity for both the amount of noise and the complexity of the cost function. All the different complexity levels have datasets containing 50 projects and 150 projects, resulting in a total of 12 potential dummy datasets that can be used for the dummy analysis. The dummy analysis is started by looking at the effect of noise on the modeling performance which is described in subsection 6.4.1 and afterward, the modeling performances of the different cost functions are presented in subsection 6.4.2. Both the parts of the dummy analysis are conducted using the optimization algorithm using the same settings.

6.4.1. Different noise levels

The first dummy analysis that is conducted is the dummy data with the easy noise level containing 50 datapoints. The results of this analysis are shown in Table 6.19, as can be seen in the table the results of this model are almost perfect already. Because of this, it is chosen not to simulate the easy noise level with 150 projects since there is, almost, no room for improvement.

Table 6.19: Algorithm results dummy analysis easy noise level and 50 projects

Dummy model noise easy 50 projects

Algorithm inputs		**Algorithm outputs**	
start_input_par	[]	**input_par**	[’Years ago’, ’AVG diameter’, ’Inhabitants’, ’Water presence’, ’Dewatering’]
n_kfold	5		
Reg_start	False		
Bagging_start	False	**Reg**	TRUE
Threshold_cor	0.1	**Lambda**	0.3
Considered_L1_wt	[0, 0.5, 1]	**L1_wt**	1
Lambda_step	0.1	**Bagging**	TRUE
n_models_bagging_step	5	**n_models_bagging**	10
R2 test AVG: 0.996			

Because of the almost perfect model performance at the easy noise level, it is chosen to skip the normal noise level and immediately go to the hard noise level to see if the model performance will decrease. The model performance of the dummy model with 50 projects and the highest noise level (hard) is shown in Table 6.20. Also, the model performance of this model is very close to perfect, this indicates that the level of noise has relatively little effect on the model performance in the chosen range of added noise. Adding extra data points for these models is therefore not required. However, since the dummy models clearly significantly outperform the real models it is very likely that the real gathered project data has higher noise levels than currently considered in the dummy analysis.

Table 6.20: Algorithm results dummy analysis hard noise level and 50 projects

Dummy model noise hard 50 projects			
Algorithm inputs		**Algorithm outputs**	
start_input_par	[]	**input_par**	['AVG diameter', 'Inhabitants', 'Building year diff med and mean', 'Water presence', 'PE_medium']
n_kfold	5		
Reg_start	False		
Bagging_start	False	**Reg**	TRUE
Threshold_cor	0.1	**Lambda**	0.1
Considered_L1_wt	[0, 0.5, 1]	**L1_wt**	1
Lambda_step	0.1	**Bagging**	TRUE
n_models_bagging_step	5	**n_models_bagging**	10
R2 test AVG: 0.969			

An interesting conclusion that can be drawn from the above two analyses is that even though the two dummy models realize almost perfect modeling scores ($R2 = 1$ is perfect score) they did not find all the right input parameters. The easy noise model found all the input parameters except the *"length open sleuf"* and instead added *"dewatering"*. The dummy model with the hard noise level also found three correct input parameters, but instead of choosing "years ago" it picked, the strongly correlated, Building year diff med and mean, and instead of using "length open sleuf" the non-related input parameter "PE_medium" was added. When comparing the found formulas, Equation 6.1, Equation 6.2, to the formula that was used to generate the cost data, Equation 6.3, it can be concluded that the ML finds the linear terms in the cost function almost perfectly. However, for the rest the two resulting dummy functions look relatively different when considering that they are able to make almost perfect predictions of the costs per meter. This is an important conclusion also for the real models because it indicates that you do not need a perfect copy of the real cost function to make accurate predictions.

$$\begin{aligned} \hat{y}_{noise\_easy} = & 86.7 + 1.17 * AVG\_diameter + 0.33 * years\_ago + 1.49 \\ & * Inhabitants + 112 * Water\_presence + 0.17 * Dewatering \end{aligned} \tag{6.1}$$

$$\begin{aligned} \hat{y}_{noise\_hard} = & 86.2 + 1.21 * AVG\_diameter + 1.28 * Building\_year\_diff\_med\_mean \\ & + 1.61 * Inhabitants + 109 * Water\_presence - 0.07 PE\_medium \end{aligned} \tag{6.2}$$

$$\begin{aligned} y_{real} = & 150 + 0.2 * AVG\_diameter^{1.3} - 4 * Ln(1 + Length\_open\_sleuf) \\ & + 0.3 * year\_ago + 1.5 * Inhabitants + 110 * Water\_presence \end{aligned} \tag{6.3}$$

6.4.2. Different cost function complexities

After looking at the different noise levels the dummy data with different cost functions is analyzed. The dummy model with the easy cost function is the same as the dummy model with the easy noise level since the easy noise level is considered for all cost functions. The results of this model, almost perfectly performing model, are presented in the previous section in Table 6.19. The first new model that is analyzed is the dummy model with the normal cost function using 50 data points. The results of this model are presented in Table 6.21, as can be seen in the table increasing the complexity of the cost function has a bigger impact on the modeling performance than adding the noise did.

Table 6.21: Algorithm results dummy analysis normal cost function and 50 projects

Dummy model cost function normal 50 projects			
Algorithm inputs		**Algorithm outputs**	
start_input_par	[ ]	**input_par**	[’Years ago’, ’Length overig’, ’Water presence’, ’Building year diff med and mean’, ’Dewatering’, ’Road Lokale weg’, ’Road Straat’, ’Water utility 2’, ’Woongebied’, ’Width road’, ’Network opr 1’]
n_kfold	5		
Reg_start	False		
Bagging_start	False	**Reg**	False
Threshold_cor	0.1	**Lambda**	[-]
Considered_L1_wt	[0, 0.5, 1]	**L1_wt**	[-]
Lambda_step	0.1	**Bagging**	True
n_models_bagging_step	5	**n_models_bagging**	10
R2 test AVG: 0.86			

To check whether increasing the number of projects would increase the modeling performance of the dummy model with the normal cost function. Now the dummy model is trained using the data set containing 150 data points instead of the 50 data points used in the previous simulation. The results of the analysis using 150 data points can be seen in . It can be concluded from the table that increasing the amount of data points did indeed improve the modeling performance of the dummy model by 0.06 from 0.86 up to 0.92.

Table 6.22: Algorithm results dummy analysis normal cost function and 150 projects

Dummy model cost function normal 150 projects			
Algorithm inputs		**Algorithm outputs**	
start_input_par	[]	**input_par**	[’Years ago’, ’other_small’, ’Tunnel’, ’AVG diameter’, ’Road Lokale weg’, ’Road Straat’, ’Water presence’ ’ Asphalt [%]’, ’Open verharding [%]’, ’Soil replacement’]
n_kfold	5		
Reg_start	False		
Bagging_start	False	**Reg**	True
Threshold_cor	0.1	**Lambda**	0.1
Considered_L1_wt	[0, 0.5, 1]	**L1_wt**	1
Lambda_step	0.1	**Bagging**	False
n_models_bagging_step	5	**n_models_bagging**	[-]
R2 test AVG: 0.92			

Finally, it is checked whether increasing the complexity of the cost function will further decrease the model performance of the dummy model. To have a fair comparison the dummy model with the hard cost function is also simulated using 150 data points. The results of this simulation are shown in Table 6.23 as can be seen indeed the model performance decreased by 0.06 compared to the normal cost function.

Table 6.23: Algorithm results dummy analysis hard cost function and 150 projects

Dummy model cost function hard 150 projects			
Algorithm inputs		**Algorithm outputs**	
start_input_par	[]	**input_par**	[’Years ago’, ’Inhabitants [#/hectare]’, ’AVG diameter’, ’Water presence’, ’Width road’, ’Total length’, ’Road Straat’, ’Woning twee-onder-een-kap’]
n_kfold	5		
Reg_start	False		
Bagging_start	False	**Reg**	True
Threshold_cor	0.1	**Lambda**	0.1
Considered_L1_wt	[0, 0.5, 1]	**L1_wt**	1
Lambda_step	0.1	**Bagging**	True
n_models_bagging_step	5	**n_models_bagging**	10
R2 test AVG: 0.86			

However, when not looking at the model performances of the dummy models but at the constructed cost functions, it can be concluded that the algorithm, again performed a lot less satisfactorily. The two dummy cost functions, shown in Equation 6.4 and Equation 6.6, are not at all similar two the two original cost function, shown in Equation 6.5 and Equation 6.7, used to generate the per meter prices. This again indicates that a decent prediction for the price per meter can be made with a function that is not similar to the actual cost function. This is a very important conclusion and should be considered when linear regression is used to predict construction prices per meter. When you are only interested in the prediction itself, using linear regression seems to be a valid solution. However, when you are more interested in the reasoning behind the prediction, which factors influence the costs the most, the dummy analysis shows that the results of a linear regression model should be used with great precaution.

$$\begin{aligned}\hat{y}_{prediction\_normal} =&987 + 0.59 * AVG\_diameter + 13 * year\_ago - 197 * Road\_Lokale\_weg\\ &- 46.9 * Other\_small + 185 * Tunnel - 83.8 * Road\_straat - 999\\ &* Percentage\_asphalt - 960 * Percentage\_open\_verharding\\ &- 115 * Soil\_replacement - 114 * Water\_presence\end{aligned} \tag{6.4}$$

$$\begin{aligned}y_{normal} =&150 + 0.2 * AVG\_diameter^{1.3} - 4 * Ln(1 + Length\_open\_sleuf)\\ &+ 0.1 * year\_ago^{1.7} + 30 * Inhabitants^{0.5} + 3.5 * Road\_overig\\ &* AVG\_maxspeed - 29 * Road\_overig + 1.5 * AVG\_maxspeed\end{aligned} \tag{6.5}$$

$$\begin{aligned}\hat{y}_{prediction\_hard} =&54.4 + 1.12 * AVG\_diameter + 0.32 * year\_ago + 1.39 * Inhabitants\\ &+ 13.3 * Width\_road + 80.0 * Water\_presence - 0.024 * Total\_length\\ &- 13.9 * Road\_straat - 33.35 * Woning\_twee\_een\_kap\end{aligned} \tag{6.6}$$

$$\begin{aligned}y_{hard} =&50 + 0.2 * AVG\_diameter^{1.3} - 4 * Ln(1 + Length\_open\_sleuf) + 0.1 * year\_ago^{1.7}\\ &+ 30 * Inhabitants^{0.5} + 3.5 * Road\_overig * AVG\_maxspeed - 29 * Road\_overig\\ &+ 1.5 * AVG\_maxspeed + \frac{800}{Percentage\_trees} + Width\_road^{1.9}\end{aligned} \tag{6.7}$$

6.5. The best performing models

In this subsection, the best performing models for predicting water and gas per meter prices are presented in a more detail. The best model is the model that has the highest average R2_test score when predicting gas or water. This means that the combined water models and the normal water models are both considered when picking the best water model.

6.5.1. The best gas model

The best gas model found in this research is the combined gas model using non-linear input transformations. This model was found using the optimization algorithm and the algorithm outputs were presented in Table 6.17. In Table 6.24 the realized $R2\_test$ scores and the p_value of the Breusch Pagan test for all k-fold splits are shown. In the table, it can be seen that even the worst-performing split (0.2862) is still significantly higher than the baseline model performance (0.16). The Breusch Pagan test, tests homoscedasticity, the 0 hypothesis of the test is that the model is homoscedastic (this is preferable see subsection 2.2.1), which means that the higher the p_value the higher the change that the model is in fact homoscedastic. Unfortunately as can be seen in the table only 2 out of the 5 splits seem to be homoscedastic. This indicates that not all the underlying assumptions of a regression model are met.

Table 6.24: Model performances (R^2) and the p_value of the Breusch Pagan test for the best model gas

kfold split	R2_test	p_value Breusch Pagan test
1	0.2865	$7.0 * 10^{-7}$
2	0.4992	$3.2 * 10^{-14}$
3	0.6218	0.48
4	0.5881	0.49
5	0.6354	$1.38 * 10^{-8}$
	R2 test AVG: 0.53	

Besides looking at the model performance, it is also interesting to look at the significance of the final input parameters. The more significant a final input is the higher the change that this input indeed has a strong influence on the price per meter of gas construction projects. Which also means that the change is higher that this input parameter influences the per meter price of DH networks. As explained in subsection 2.2.2, a measure of significance for a single input parameter is the p_value of a t-test. The average p_values for the t-test of all the input parameters in the best performing gas model are presented in Table 6.25. The p_value is an indication of the change that a certain parameters is included in the model based on random change and is therefor useless for cost predictions. A lower p_value means a lower change (of random including) resulting in a more significant input parameter.

Table 6.25: Average t-statistic (p_values), theta and unit of input parameters best performing gas model

Input parameters	AVG t-statistic	theta	unit	Input parameter	AVG t-statistic	theta	unit
Other_medium	0.53	-2.1	[m]	Pollution category	0.13	111	[-]
PE_medium	0.46	-0.058	[m]	PVC_big	0.075	-0.0771	[m]
AVG maxspeed	0.45	0.06	[km/h]	AVG diameter	0.006	1.58	[mm]
Years ago	0.44	0.065	[#]	1/(Total length)	0.0057	19859	[1/m]
PVC_medium	0.32	0.04	[m]	Network opr 1	0.00092	-276	[-]
Woning vrijstaand	0.13	-84.8	[-]				

Maybe the most important result of the gas analysis is the final cost function that is constructed to predict per meter prices of gas projects. This cost function is shown in Equation 6.8. This cost function is used to predict all the costs of the 52 gas projects. The results of these predictions (y_hat) together with the actual cost (y_real) can be seen in Figure 6.3. In the graph, all the projects are ordered based on their construction costs, and the project number (after ordering) is shown on the x-axis. This means that the pattern that could be identified when connecting all the dots in the graph does not have any meaning and just indicates the distribution of the construction costs for the gathered project data.

$$
\begin{aligned}
y_{gas} =& 301 + 1.58 * AVG\_diameter - 276 * Network\_opr\_1 + \frac{19859}{Total\_length} \\
& + 111 * Pollution\_category - 0.0771 * PVC\_big - 84.8 * Woning\_vrijstaand \\
& - 0.04 * PVC\_medium - 0.058 * PE\_medium + 0.065 * Years\_ago \\
& - 2.1 * other\_medium + 0.06 * AVG\_maxspeed
\end{aligned}
\tag{6.8}
$$

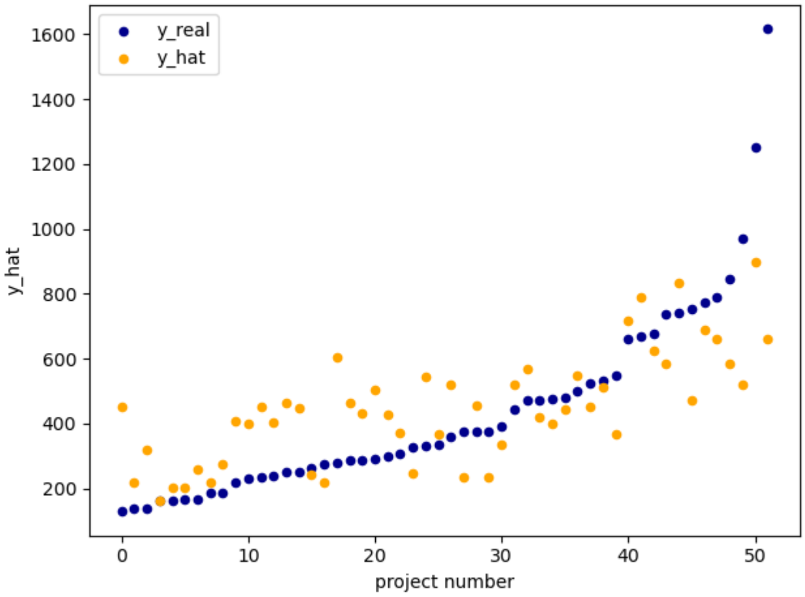


Figure 6.3: Gas construction cost predictions using best performing model (y_hat) and actual construction costs for gas projects (y_real)

Finally, short explanations of the input parameters used in the best performing gas model are given in Table 6.26. A more elaborate description of the data gathering and data preprocessing can be found in section 4.2 and section 5.1 respectively.

Table 6.26: Explanation of input parameters used in the best performing gas model

Input parameters	Explanation
Other_medium	Number of meters of a material that is not PVC or PE in with a diameter bigger than 75 and smaller than 125. Also, see Table 5.2 for all the material and diameter categories.
	-
PE_medium	Number of meters of PE with a diameter bigger than 75mm and smaller than 125mm in the project. Also, see Table 5.2 for all the material and diameter categories.
	-
AVG maxspeed	The average speed limit in the project area. The weighted average of all the streets is conducted the speed limit data is found in OpenStreetMap.
	-
Years ago	The average amount of years ago that a building is constructed. From all the building in the project area, the building year is inputted from BAG and subtracted from 2020.
	-
PVC_medium	Number of meters of PVC with a diameter bigger than 75mm and smaller than 125mm in the project. Also, see Table 5.2 for all the material and diameter categories.
	-
Woning vrijstaand (Detached house)	This boolean indicates whether most of the houses in the project area are detached houses. The used house types are imported from an ARCGIS layer containing house types.
	-
Pollution category	Ground polution level is either 0,1 or 2 see \autoref{tab:pollution cat to num}. For now, the pollution levels are gathered in the data gathering spreadsheet but it is expected that a national database containing ground pollution levels will become available.
	-
PVC_big	Number of meters of PVC with a diameter bigger than 125. Also, see Table 5.2 for all the material and diameter categories.
	-
AVG diameter	The weighted average diameter of all pipes in the project. Calculated based on the number of meters for every diameter.
	-
1/(Total length)	one divided by the total number of meters that are constructed in the project. This non-linear input parameter transformation is calculated using the total project length. Which is gathered in the data gathering spreadsheet.
	-
Network opr1	Boolean to indicate that historical data was received from network opr 1.

6.5.2. The best water model

In Table 6.27 the model performances and the p_value of the Breusch Pagan test of the different splits for the best performing water model are shown. As can be seen the worst-performing split (0.107) still significantly outperforms the baseline model (0) and three out of the splits seem to be homoscedastic.

Table 6.27: Model performances (R^2) and the p_value of the Breusch Pagan test for the best model water

kfold split	R2_test	p_value Breusch Pagan test
1	0.338	0.000397
2	0.442	0.00116
3	0.478	1.28
4	0.741	0.000287
5	0.460	5.82
6	0.107	1.17
	R2 test AVG: 0.43	

In Table 6.28, again, the average p_values of the t-test for all the parameters included in the best performing water model are depicted. When looking at the p_value of the best performing water model it can be seen that as for the gas model, the AVG diameter, 1/(Total length) and the pollution category are again relatively significant. Indicating that these three input parameters are the most promising for predicting construction costs of infrastructures that are similar to water and gas infrastructures and for that reason are also likely to have added value for predicting DH prices.

Table 6.28: Average t-statistic (p_values), theta and unit of input parameters best performing water model

Input parameters	AVG t-statistic	theta	unit	Input parameter	AVG t-statistic	theta	unit
PE_small	0.294	0.087	[m]	Pollution category	77	[-]	0.174
Water utility 1	0.268	62	[-]	1/(Total length)	16166	[1/m]	0.031
Water utility 2	0.252	-53	[-]	AVG diameter	1.6	[mm]	0.0084

In Equation 6.9, the constructed cost function for predicting prices per meter of water projects is shown. Again also the predictions that result from implementing this cost function to predict the construction cost of the 68 water projects can be seen in Figure 6.4.

$$\begin{aligned} y_{water} = & -31 + 1.6 * AVG\_diameter + \frac{16166}{Total\_length} - 53 * Water\_utility\_2 \\ & + 77 * Pollution\_category + 0.087 * PE\_small + 62 * Water\_utility\_1 \end{aligned} \tag{6.9}$$

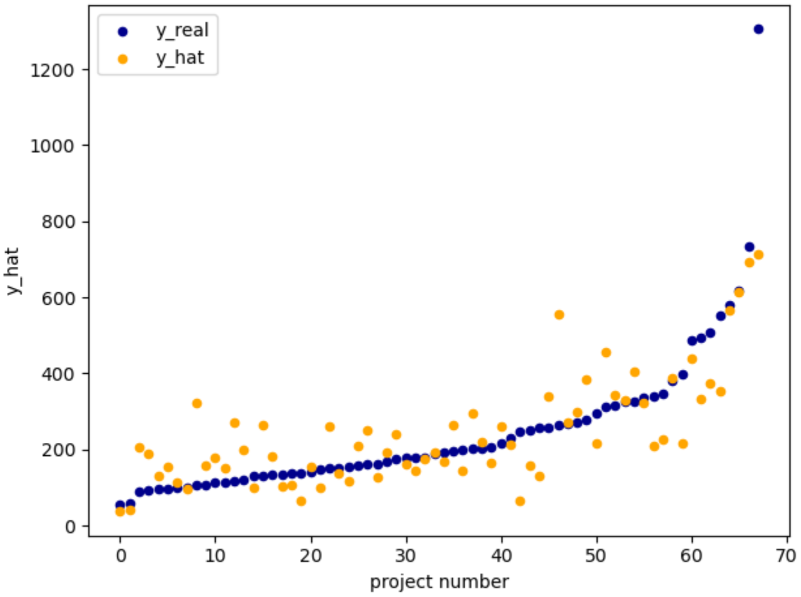


Figure 6.4: Water predictions using best performing model (y_hat) and actual construction costs for water projects (y_real)

Finally, in Table 6.29 again short explanations on the used input parameters for the best performing water model are given.

Table 6.29: Explanation of input parameters used in the best performing water model

Input parameters	Explanation
PE_small	Number of meters of PE with a diameter smaller than 75mm. Also, see Table 5.2 for all the material and diameter categories.
	-
Water utility 1/2	Boolean to indicate that historical data was received from water utility 1 and water utility 2 respectively.
	-
Pollution cateogry	See Table 6.26
	-
1/(Total length)	See Table 6.26
	-
AVG diameter	See Table 6.26

6.6. What does this mean for Heat

The main goal of this research is to contribute to DH network cost predictions. Therefore, in this section, the potential added value of this research for DH predictions is discussed. The first and probably most important conclusion is that the modeling approach used in this research shows promising results for gas and water and is therefore likely to also have decent results for DH predictions. Both the best performing water and gas models (R2_score: 0.53 and 0.43) significantly outperformed the baseline models (R2_score: 0.16 and 0). Indicating that using linear regression with more inputs than solely the diameter can improve the prediction accuracy and thereby decrease the bandwidth of cost predictions. When converting the R2_score of the best performing water and gas model into a reduction in variance and using this variance reduction to calculate the band with reduction, see Equation 6.10, it was found that the water and gas model both reduce the size of the bandwidth with approximately 25%. Of course, there is no guarantee that the same 25% reduction can be realized by DH, but the perspective is very promising.

$$Bandwith_{95\%} = 2 * \sigma \tag{6.10a}$$

$$\sigma = \sqrt{VAR} \tag{6.10b}$$

$$Reduction\_Bandwith = 1 - \sqrt{\frac{VAR_{new}}{VAR_{old}}} \tag{6.10c}$$

$$Reduction\_Bandwith\_Gas = 1 - \sqrt{\frac{1-0.53}{1-0.16}} = \mathbf{0.252} \tag{6.10d}$$

$$Reduction\_Bandwith\_Water = 1 - \sqrt{\frac{1-0.43}{1-0}} = \mathbf{0.245} \tag{6.10e}$$

A second promising contribution of this research is the developed optimization algorithm. For both the water and gas model, the resulting model from the optimization algorithm outperformed the best model found by changing inputs and modeling settings by hand. This optimization algorithm is developed in such a way that it can also be used for DH model optimization if DH data is available in the future.

Thirdly also the gathered historical project data for water and gas projects has the potential to be of added value for DH predictions. Based on the results of this analysis, a rough indication can be made, about the change of increasing the model performance, when the data for multiple (similar) infrastructures is merged. Based on this analysis there is approximately a 50% change that including historical data from similar infrastructures in the training process of a linear regression model improves the modeling performance. Since including water data in the training data set of the gas model improved the gas model performance but including gas data in the training set of the water model did not improve the performance. Therefore, it could be that adding the water and or gas data to a training set containing DH data also improves the modeling performance. This of-course is not certain but very promising to validate in follow-up research.

Finally, two important conclusions can be drawn based on the dummy analysis. First, the dummy analysis showed that interpreting the results of a linear regression model should be done with caution. When one is interested in the parameters that influence the cost instead of the actual cost prediction. It is likely that this phenomenon is also applicable to a future heat model. Secondly, the cost functions for gas and water models are probably quite complicated. Because the modeling performance of the gas and water models is significantly lower than the dummy model performance. Accurately predicting these complicated cost functions with linear regression is therefore expected to work significantly better when more data points are used. When more data points are used, also more input parameters can be used in a statistically significant way, remember the rule of thumb discussed in section 5.3. It is likely that also the DH cost function is complicated and that using more data points has the potential to significantly improve the model performance.

7 Discussion

In this research linear regression is applied to model the construction cost per meter of water and natural gas networks. The goal of the research is to find relations between this construction cost and surrounding parameters which can be used in future research to model district heating costs in the built environment. However, some important limitations should be taken into account when using the results from this research. These limitations are discussed in this chapter.

First, in section 7.1 it is discussed how the results from this research can be used for district heating purposes specifically. In this subsection also some critical differences between water and gas and DH networks are discussed which should be taken into account when the results from this research are used for DH predictions. Secondly, section 7.2, reviews important factors that are likely to influence the construction cost of infrastructures in the built environment but are not included in the modeling process of this research. Afterward, unexpected modeling results with potential explanations are given in section 7.3. Then some critical nodes resulting from the conducted dummy analysis are elaborated on in section 7.4. Finally, in section 7.5 two validations of the model are discussed. First, in subsection 7.5.1 a sensitivity analysis on the chosen buffer size for selecting surrounding parameters from the GIS databases is presented. Second, in subsection 7.5.2 the Mean Squared Error (MSE) of the best models is compared to the baseline models to validate the increase in modeling performance that is found when looking at the R^2 values.

7.1. Using research results for district heating predictions

The main focus of this research is surrounding based cost prediction for DH networks. However, since no DH data is available the modeling part of the research was only based on gas and water projects. Fortunately, the results of the water and gas models can be of added value for DH predictions in two ways.

First of all the methodology applied in this research has proved its worth and can very easily be applied on DH historical project data. Keeping in mind that the resulting models for both water and gas significantly outperform the baseline model gives reason to believe that similar results are realistic for DH networks when sufficient historical project data is available. It however important to realize that, because of the lack of available data for DH projects as discussed in section 1.2 and the fact that commercial (heating) companies are less likely to cooperate in the time-consuming data sharing process, gathering sufficient data only considering DH projects is a very challenging task. For this sole reason, this thesis was initiated to identify if water and gas data can also be used as an (extra) input for a DH cost model.

Using the data and conclusions (important surrounding parameters) from the water and gas models is the second way in which this research can be a contribution to DH cost predictions. Including promising (surrounding) parameters like pollution category and 1/(total length) (no surrounding parameter) in heat, prediction models is promising. Also when considering that using the water data to train the gas model improved the gas model performance significantly, it can be argued that this might also be true for the modeling performance of a DH model using gas and or water training data. However, it is of course not certain that using the water and gas data is of added value for a heat model since using the gas data for water did not increase the modeling performance. Because of the potential added value, it is recommended that in future DH research the water and gas data is used and it is validated whether the model performance is increased.

There are however some important differences between water and gas, and heat infrastructures that might cause limitations when combining the data of the three different infrastructures. These limitations should be taken into account when using the water and gas data to train a DH model.

The first and probably most important difference between water and gas, and DH networks is the location of the pipes in the street profile. Based on the interviews with experts it was concluded that the most common location for gas and water networks is below the sidewalks. Sidewalks are a preferred location because they generally do not consist of asphalt and are closer to the houses. The problem however is that sidewalks are getting full and that it is unlikely that there is room left for the big DH pipes. Meaning that probably most of the future DH networks need to be constructed below the street instead of the sidewalks. Constructing a network below the street is more expensive because generally more traffic measures and asphalt cutting is required. To verify this hypothesis that originated in an interview with Warmtestad Groningen an interview with an urban planner from the municipality of Amsterdam was arranged. In this interview, it became clear, that at least in Amsterdam, it is indeed very likely that future DH networks should be constructed below the street. This difference in location could for example lead to misinterpretations, of the extra water and gas training data, for surrounding factors like the road coverage. The road coverage is less likely to influence the costs of water and gas networks but it is probably a very significant factor in the cost of DH networks.

Besides the location in the street, another important difference that is likely to influence the modeling performance when the data sets are combined is the average diameter. The average diameter is an important input parameter for all infrastructures, however the average diameter for water and gas networks are a lot more similar than the average diameter for a DH network. This difference could cause an offset in the linear regression model which should be compensated.

A third important difference is that the considered water and gas projects are replacement projects whereas in the heat projects a new infrastructure is established. This results in a couple of key differences, first of during the construction process the connected households still need access to water and gas. Meaning that either they have to construct a small peace every day, so that the water or gas is only cut off for a single day, for a couple of households at the time. Or they first have to construct the new infrastructure before the old pipes can be removed. Removing these old pipes is the second difference between replacement projects and new projects. Because of this difference, in this research, it was tried to extract the removing costs from the model. However, it turned out that the extra costs for removing the old infrastructure were hard to reproduce from the available project data. Therefore, these costs, in the end, were not extracted in the final model. A final potential difference between a replacement project and a new project is the extra logistics and information services required to connect new houses to new infrastructure.

Furthermore, another important difference regards the contracts and the already made deals in the infrastructure constructing sector. Most water and gas utility companies have already settled deals with other infrastructures on how they divide the cost when they are working together. If heat is new in this dynamic, and also when they need to start working as soon as possible, they do not have a great negotiation position. This could potentially lead to a bad deal and therefor higher per meter prices for DH network construction.

Finally, the, most common, material type is also an important aspect to consider. Older, higher temperature, DH networks were generally constructed using steel pipes. The welding process of these steel pipes requires more space (deeper and wider trench) and is more time consuming than the construction of plastic (PVC / PE) pipes that are most commonly used for water and gas distribution networks. This difference however is probably becoming less relevant, since new, lower temperature, DH networks are more likely to be constructed using plastic-based pipes (PVC/PE). When this happens the predicting potential of water and gas projects compared to old DH projects using steels pipes is actually becoming more relevant.

7.2. Factors that might be important but are not included in the models

Before the data gathering process started first interviews were conducted to identify important factors that can influence the cost of constructing (fluid) infrastructures in the built environment. Not all the factors that were deemed relevant were also included in the final models. Also for some factors that were included, the used database does not seem to provide sufficient data. In this subsection, all the limitations on factors that might influence the costs but are not (sufficiently) included are discussed. The fact that these factors are not included in the model means that it is not certain if indeed all these factors have a significant influence on the cost. However, it is very likely that at least a few of the mentioned factors are important and should be added in future research when new data becomes available.

The first category of factors that were not included in the model consist of surrounding input parameters of which no national database was found, or in the case of the klic database was not available for this research.

The factors are:

- Amount of other infrastructures in the ground (klic)
- Not exploded explosives (NGE in dutch)
- Traffic density
- Metro line crossings
- Private landowners

For the not exploded explosives a smaller (GIS) database only covering parts of the Netherlands is found which can be included when in the future a case study is performed on a specific area in the Netherlands. For the number of other infrastructures in the subsurface, the klic databases exist, however unfortunately the data available in the klic database is highly confidential and was therefore not available for this research. The private landowners' data was considered to be only significant for transport pipes outside of the built environment and is for that reason excluded from the research.

The second category of factors consists of surrounding parameters that were included in the final modeling process. But when including the data some questions arose about the use trustworthiness of the data. The first database in this category is the database containing the "trefkans" (probabilty of) archaeology in The Netherlands. When using this database it was realized that the database contains the value "unknown probability" for big parts of the country (especially in city centers). This resulted in the fact that for a significant amount of the historical projects used in the research the corresponding archaeology probability was unknown and therefore not use full as model input. The next database that did not seem to provide the added value that was sought for is the average groundwater level database. The reasoning behind using this database is that when the groundwater level is high the chances of dewatering increase and therefore the costs increase. However to validate that the average groundwater was indeed related to dewatering, in the data gathering spreadsheet a yes/no question was included asking whether dewatering was applied. When looking at the correlation between the average groundwater level and the dewatered projects for both the water and gas projects it was concluded that the average groundwater level database was not correlated. The correlations between dewatering and the average groundwater level are 0.038 and 0.001 for water and gas projects respectively. A similar validation is conducted for the ground-type database, when the ground type consists of clay the ground is likely to be replaced because clay will sag. However, the correlation between whether ground replacement is conducted and whether the name clay was found in the ground type database was also very weak.

The final category of factors that could influence the cost but that are not included are factors that are not easily captured in databases and therefore also not easily implemented in ML models. The first example of such a factor is the tightness of the contractors market. The price of a contractor, and thus of the total project, is strongly dependent on supply and demand, this supply and demand changes over time and possibly from location to location. Because of this reason, it is chosen to collect projects that are as recent as possible as input for this research. However, even in a period of a couple of years (oldest project is 2018), the contractor prices can change, and even more importantly the contractor prices will keep changing in the future. Meaning that constructed cost parameters will lose accuracy over time. Furthermore, some "hidden" surrounding parameters could potentially influence the cost but are very location specific and hard to identify for the entire country. Three examples of these that were identified in expert interviews are:

- In Amsterdam there are strict rules on when you are allowed to work and when you're not allowed to work, also in a lot of projects, it is mandatory to work together. These rules could potentially influence the cost by increasing the complexity of the planning of the project.
- In Amsterdam the municipality has to be hired to close the road after a project. This cannot be done by the contractor himself. The prices the municipality charges for this are a lot higher than the "normal" market price asked by regular contractors.
- In Limburg because of hills they sometimes need higher pressure (Steel) pipes, which are more expensive but last a lot longer.

7.3. Unexpected modelling results

Not all the modeling results presented in chapter 6 were in line with the expected results. Some of the most important unexpected modeling results together with a potential reasoning behind it are given in this subsection.

The first example of a result that was not expected, is the fact that the amount of other infrastructures that are "meegekoppeld" in a certain project is not used in any of the well-performing models. This would indicate that this input parameter is not significant for the price per meter when it would be expected to be of significant influence. A potential reason for this could be that a relatively big part of the projects that were gathered were solo projects (67%). Another reason would be that the amount of other infrastructures constructed in the project are less important than expected beforehand.

The second counter-intuitive result is the fact that the gas model outperforms the water model. Since the water model has more data points it would be expected that the water model outperforms the gas model. The fact that the water model performs less can have a lot of reasons, but one that stands out is the fact that the collected data is not equally separated over the different utility companies providing the water data. This increases the chance that all or a significant amount of the projects from a certain water utility end up in the same set (training set/testing set) when this happens this is likely to reduce the model performance.

A third result which is not anticipated beforehand is the fact that the building year does not seem to have a big impact on the costs. When almost all the experts specifically pointed out that working in a historical city center is the most expensive project type. The fact that this is not found in the trained models can be because not that many projects from historical city centers are included in the analysis.

The fourth results that is a bit curious, is the fact that including water data in the gas model increases the modeling performance (significantly). Whereas, including gas data in the water model decreases the performance (significantly). It would be expected that if the infrastructures are indeed similar combining the data from both infrastructures increases the modeling performances of both models. The possible explanation for the fact that the water model performance is not increased by adding gas data is the number of projects that are considered. Adding the 68 water projects to the gas training data set which contains 42 or 41 data points might be more beneficial than adding the 52 gas projects to the water training data set which contains 56 or 57 data points.

A final model results which was expected but is undesired is the fact that the utility that provided the data is a relatively significant input for predicting the cost. This can have three reasons. Either a certain utility operates in a hard area, when this is the case the utility should not be used as input but the responsible surrounding parameters should explain the effect. Or some utility companies are more expensive because they are less efficient or have higher profit margins on projects. Finally, it can also be because some utilities made different assumptions on which cost to include and which cost to exclude when providing historical project data. Since it is not known which of the three above mentioned reasons (combination can also be true) causes the significance of the data providing utilities, it is recommended to keep including the data providing companies in the first iterations of future research. It is however promising to further look into the three above-mentioned reasons to see if the most important reason can be identified.

7.4. The dummy analysis

In this subsection specifically, the results of the dummy analysis are discussed. The dummy analysis has its own subsection in the discussion because the results of the dummy analysis lead to two important discussion points that are discussed below.

The first discussion point has to do with the way the ML model performance is scored and how the results of the analysis can and should be interpreted. When looking at the model performances of the dummy models it can be concluded that the dummy models perform very well (high R2_test scores). However, when looking at the resulting cost functions and comparing these to the "real" cost functions used to generate the data it can be seen that these differ quite significantly. This leads to the question, does the dummy model actually perform well? To answer this question it is very important to know what you're interested in. Generally speaking, a ML model can be used with two different intentions. The first prediction is to predict, a (dependent) variable, and the second is to explain the behavior of a variable. When one is interested in the first function, predicting, and is only interested in the final result (cost per meter), then the dummy model, and therefore also the water and gas model, is likely to perform well, since it can accurately predict this cost. However, when one is interested in the important factors that influence the cost (the model is used for explaining purposes) the dummy model performs less good, since some of the factors in the cost formula or not supposed to be there. For these factors, it is not certain whether they influence the cost (because they are correlated to factors included in the formula) or are included based on random chance. A nice example of the former is the factor $building\_year\_diff\_med\_mean$ which is included in the dummy cost formula of the hard noise model shown in Equation 6.2. Even though the $building\_year\_diff\_med\_mean$ itself is not included in the original cost function the input parameter $years_ago$ is. Since these two input parameters are correlated significantly this means that the input parameter $building\_year\_diff\_med\_mean$ is a decent choice for predicting the costs. However, in the dummy cost functions, shown in Equation 6.4 and Equation 6.6, some input parameters in the final cost function are very hard to relate to input parameters in the "real" cost functions and are therefore likely added based on random chance. A good indication of which parameters are more likely to be related to cost and which parameters are more likely to be added due to random chance is the p_value of the t statistic. These p_values of the best performing water and gas models are given in Table 6.28 and Table 6.25 respectively. Whether or not it is a problem that some parameters in the final cost functions are likely added to due random change is up to the person using the model. However, it is a very important conclusion that should at least be considered by people using the model results to identify potentially important surrounding parameters.

The second point of discussion is the big gap in model performance between the dummy models and the best water and gas models. This big gap could indicate that the cost function of the price per meter of a (fluid) infrastructure in the built environment is to complicated to model properly with a linear regression model. However, it could also mean that indeed the cost function is more complicated but that when using more data points this cost function is perfectly predictable using linear regression. Looking at other papers that implemented ML for cost prediction it can be concluded that on average more data points were used than were available in this study. Especially when also keeping in mind the number of potential inputs considered in this research adding more data points has a big potential of increasing the modeling performance. Since, as already stated in section 5.3, ideally per input parameter 15 till 20 data points are used. For the best performing models for water and gas respectively 6 and 12 input parameters were used meaning that this ideal situation is not realized.

7.5. Validation

In this section of the discussion, two critical decisions that were made in this research and were deemed critical are validated. First, in subsection 7.5.1 a sensitivity analysis on the chosen buffer size is conducted. Secondly, another scoring criteria (MSE) is used to score the best performing models to validate the main scoring criteria (R^2) in subsection 7.5.2.

7.5.1. Sensitivity analysis final surrounding parameters

In this subsection, a sensitivity analysis on the buffer size is conducted. As explained in subsection 4.2.3, the buffer size, that is used to clip the GIS databases, in this research is set to 30m. This decision is made based on looking at ten different projects by hand and an original sensitivity analysis that checked the effect of the buffer size on 31 (water) projects. Based on this original sensitivity analysis it was concluded that most surrounding parameters did not seem to change significantly when the buffer size was changed. However, since, this sensitivity analysis was only conducted for 31 projects it is important to check that for the relevant surrounding parameters the buffer size indeed does not influence the parameters significantly when all projects are considered.

First, let us take a look at the best performing water model. In the best performing water model, see subsection 6.5.2, the only surrounding parameter that is included is ground pollution. Since ground pollution levels are gathered through the data gathering spreadsheet instead of a GIS database no sensitivity analysis is required for the water model. Because non of the selected input parameters is influenced by the buffer size.

Now let us consider the best performing gas model. In the best performing gas model three surrounding parameters are included namely: *AVG maxspeed, Years ago* and *Woning vrijstaand*. For these three surrounding parameters, a sensitivity analysis based on all (gas) projects is conducted. The only difference is that this time instead of the 4 considered buffer size only 3 buffer sizes are compared to save computational time. The buffer size that is excluded compared to the original sensitivity analysis is 35m. Since this buffer size is in the middle of the sample it is assumed that removing this buffer size has a marginal impact on the analysis. In the sensitivity analysis again the numerical surrounding parameters and categorical surrounding parameters are evaluated differently.

For the categorical surrounding parameter, *Woning vrijstaand*, it is checked for how many of the 52 gas projects the most common building type changed when the buffer size is changed from 30 to both 25 and 45. It was concluded that for non of the 52 gas projects changing the buffer size changed the most common building type. Meaning that for the input parameter *Woning vrijstaand* it literally does not matter if the buffer size is 25, 30, or 45 meters.

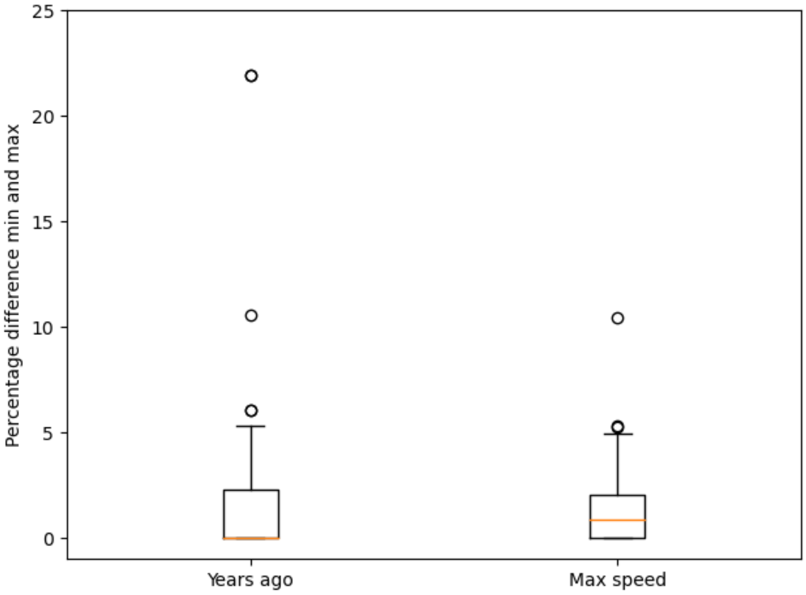


Figure 7.1: What is the difference, in percentage, between the maximum and minimum value of the 2 numerical surrounding parameters in the final gas model for the 3 considered buffer sizes (25m,30m,45m)

For the numerical surrounding parameters, *AVG maxspeed* and *Years ago*, again the formula depicted in Equation 4.1, is used to calculate the percentage difference between the minimum and maximum value out of the three different buffer sizes for all the 52 gas projects. Two boxplots presenting the results of this analysis for all the gas projects for both surrounding parameters can be seen in Figure 7.1. As can be seen only for a very limited amount of projects changing the buffer size significantly influenced the two surrounding parameters. This again indicates that the buffer size is not likely to be a critical factor in the modeling approach used in this thesis.

7.5.2. Look at MSE scoring criteria to validate increasing model performance

As serious increase in model performance is realized when looking at main scoring criteria (R^2) applied in this research. However it is important to validate that this increase of the R^2 value is indeed because the model is significantly better in prediction construction costs than the baseline model. As mentioned in section 5.2 it is also possible than a certain model performs really well for a certain scoring criteria based on random change. To validate that this is not the case for this research also the MSE of the best models is compared to the MSE of the baseline model.

Table 7.1: MSE test scores of the baseline and best performing models

Model type	**MSE test score**						
	AVG	split 1	split 2	split 3	split 4	split 5	split 6
Baseline Gas	**67844**	154854	72876	36494	41385	33622	
Best model Gas	**44448**	115864	53889	21240	19443	11802	
Baseline Water	**29602**	8202	68574	27440	5604	40954	26837
Best model Water	**19182**	3048	59871	17970	1481	15110	17611

In Table 7.1 the MSE of the best performing models and the baseline models for both water and gas are presented. As can be seen in the table the MSE decreased (significantly) for all the different splits (from k-fold splitting) for both the best performing water and gas model. On average the MSE decreased by **34%** for the gas model and with **35%** for the water model. This very serious reduction indicates that indeed the developed water and gas model can predict the construction costs significantly better than the baseline model. In fact the reduction found in the MSE, (34% and 35%) is larger than the bandwidth reductions (25%)that are calculated based on the R^2 score in section 6.5. Meaning that based on this validation it is more likely that the R^2 underestimated the model performance instead of overestimating the model performance. This is a very positive result emphasizing the strength of the used modeling approach.

8 Conclusion and Recommendations

In this concluding chapter of the report, the most important conclusions and some recommendations for future research are presented. First in section 8.1 the conclusions are presented and subsequently in section 8.2 the recommendations are given.

8.1. Conclusion

In this conclusion section, the research question and the corresponding subquestions that were introduced in section 1.4 are answered. First of the answers to the six separate subquestions are given. Afterward, the most important parts of these answers are merged to answer the main research question at the end of this section.

1. How do current energy transition models calculate construction cost for district heating networks? In chapter 2 a state of the art analysis of cost calculation tools was conducted. In section 2.1 specifically energy transition models calculating ,among other things, construction costs for DH networks were examined. In this analysis, 13 different energy transition model were studied with a specific focus on how these models calculate the construction cost of DH networks. It was concluded that there is significant room for improvement regarding the detail level of the cost calculations. Specifically, three important conclusions were drawn from this state of the art analysis. First, the bandwidth of the currently available models is relatively high, meaning that there is a high level of uncertainty concerning the prediction of construction costs (per meter) of DH networks. Second, it is likely that a significant part of this bandwidth results from the fact that no or hardly any surrounding parameters are considered when predicting the construction costs. Whereas it would be expected that surrounding parameters do have a significant influence on the construction costs. The final important conclusion is that most (10/13) studied models offer the possibility to enter own cost key figures. In these models it is possible to introduce your own expected costs (price per meter) as an input, that will be used for costs calculations. This means that generating more detailed cost key figures, considering surrounding parameters, can potentially improve the modeling performance of most of the studied energy transition models. This will be of direct added value to the Dutch energy transition in the built environment because these models are (meant to be) used by municipalities that are responsible for shaping this transition.

2. Which modeling approach should be used to develop a model that calculates the construction costs of pipe infrastructures in the built environment? In section 2.2 a literature review on different cost modeling techniques is conducted with the primary focus on ML approaches for costs prediction. Characteristics of different ML approaches where compared, and literature was studied in which ML was applied for similar, cost prediction, purposes. Based on this literature review it was chosen to use linear regression to develop a model to predict construction costs for this research. It was chosen to use linear regression because:

- Linear regression is the most common ML approach to predict cost for similar cost prediction projects for DH and similar infrastructures.
- When data sets are small linear regression seems to work (at least relatively) better [56].

- The results of a Linear regression model are more transparent and therefore easier to interpreted and compare with the results of other (linear regression) models, which is important for this research since a comparison between construction costs of different infrastructures is required.

3. Which infrastructures are similar enough to district heating networks to include in a combined cost modeling approach and what are the main similarities and differences between these infrastructures and DH networks? Based on interviews with domain experts two infrastructures were considered similar enough to DH to be used as extra inputs for construction costs predictions of DH networks. The considered infrastructures are the drinking water and natural gas networks. The similarities and differences between the three (water, gas, heat) considered infrastructures are presented in section 4.1. The five most important differences that should be considered when using water and gas data for DH predictions are:

- The location in the street profile, water and gas networks are commonly placed below the sidewalks whereas new heating networks are likely to be placed below the street.
- The average diameter, which is the most important cost prediction parameter, is significantly bigger for DH networks compared to water and gas infrastructures.
- The considered projects for water and gas are replacement projects. While in the DH projects a new infrastructure is constructed. When replacing infrastructures the connected households can not be disconnected from water and or gas for too long and the old pipes need to be removed. Both are potentially cost-increasing factors that should be considered when comparing the construction costs.
- Most water and gas utility companies already have settled deals regarding how to divide costs when working together with different infrastructures. Heating companies are new in this dynamic, and because they need to start working as soon as possible they are likely to have a weak negotiation position.
- The welding of steel (DH) pipes requires more space (deeper and wider trench) and is more time consuming than the construction of plastic (PVC / PE) pipes that are most commonly used for water and gas distribution networks. It is however important to keep in mind that plastic pipes are also becoming the standard in DH networks when the temperatures are decreasing and steel pipes are no longer necessary.

4. Which surrounding parameters are potentially related to the construction costs of DH networks and similar infrastructures and where can (national) GIS databases containing this data be found? Based on approximately 40 interviews with domain experts, see Table B.1 for an overview of all the conducted interviews, and discussion with colleagues at Deltares a list with potentially important surrounding parameters was drafted. This list, containing 25 surrounding parameters, can be seen in Table 4.3. Unfortunately not for all 25 parameters a corresponding GIS database was found. For some parameters no GIS data was available and for other parameters, the GIS data was not (yet) available on a national scale. In the end 5 out of the 25 parameters where excluded, leaving 20 surrounding parameters that were used as potential inputs for the ML model.

5. Which surrounding parameters have the biggest influence on the construction cost of pipe networks in the built environment? Looking at the modeling results presented in chapter 6 some promising surrounding parameters for cost prediction of (fluid) infrastructures in the built environment can be identified. The most promising surrounding parameter, that is included in both the best performing water and gas model, is the **pollution category**. Logically a higher pollution category led to higher construction costs. As can be seen in Table 5.1 the model input for the pollution category was 0, 1, or 2. Corresponding to no or light, medium and severe pollution. These levels correspond to the colors (blue, orange, and red or black) from the crow 400 standard defined by Liander. This standard is applied by default for ground quality research in The Netherlands. At this point no national database containing these pollution levels is available, and the pollution levels used in the analysis were gathered using the data gathering spreadsheet. However, local databases containing data on ground pollution levels, are available at some municipalities, provinces, and network operators. It is expected that a national database containing an indication of ground pollution levels for the entire country will become available in the near future.

Besides the pollution category, other surrounding parameters that were included in the best performing model and are therefore high potential surrounding parameters are the category value for housing types

"Woning vrijstaand", the average amount of "Years ago" that buildings in the project area are constructed and the "AVG maxspeed" for all the roads in the project area.

Furthermore, two surrounding parameters are not included in the best performing models but do show potential for cost modeling. This potential results from the fact that they have relatively high correlations with the price per meter and when these surrounding parameters were added, in the identification phase of the research (by hand analysis), they significantly improved the model performance. The two surrounding parameters, that are strongly correlated to each other and are likely to have a very similar influence on the costs, are the building footprint and the number of inhabitants per hectare. The fact that they are strongly correlated means that adding both of them to a cost function is not very sensible since they are likely to compete for the same part of the costs. For most projects the number of inhabitants and building footprint are very similar, the main difference occurs when flats are present in the project area. Since there are not enough projects in the training data set, that meet this specific situation, it needs to be validated which surrounding parameter performs best in project areas with flats.

Finally, **1/(total length)** is also a very promising factor to include in a construction costs per meter analysis. Even though this factor is no surrounding parameter and therefore not the main interest of this research. It is strongly recommended to include this input parameter because it seems to be of more added value than the most promising surrounding parameters for both the water and gas model. When looking at the 1/(Total length) graph (Figure 6.1a) in subsection 6.1.3 again, it can be concluded that gas projects (same is true for water) with a low amount of meters are significantly more expensive per meter. This is logical since the startup project cost for small projects are relatively a lot higher compared to large projects. Including 1/(Total length) as input for a cost prediction model compensates for this effect and therefore improves the modeling performance.

6. How can historical project data and resulting models for similar infrastructures best be used to improve a construction cost model for district heating networks? In the beginning of this report, in section 1.3, three ways that this research could be of added value for heating predictions were identified. three ways that this research could be of added value for heating predictions were identified. The three different contributions were: to act as a proof of concept for the modeling approach, to identify the most important surrounding parameters, and to provide potential extra training data. In this concluding part of the report, some concluding remarks are made as to the potential contribution of this research on these three aspects.

The first potential contribution, the proof of concept of the modeling approach, was achieved very successfully. Both the best performing water and gas models (R2_score: 0.53 and 0.43) significantly outperformed the baseline models (R2_score: 0.16 and -0.17). Indicating that using linear regression with more inputs than solely the diameter can improve the prediction accuracy and thereby decreasing the bandwidth of cost predictions. The fact that for both water and gas such promising results were achieved strongly indicates that the same is possible for a DH model. As discussed in section 6.6 both the best performing water and gas model reduced the prediction bandwidth with **25%**. This significant reduction is very promising, for DH as well, especially when considering the relatively small amount of data points used to realize this reduction.

For the second contribution, it is harder to identify at this moment whether it will indeed be of added value for heating predictions. It should be validated weather the most important surrounding parameters that are found in this research, also are applicable for DH networks. However, it is very likely that at least some of the surrounding parameters that are gathered in this research are useful for predicting DH construction costs. Meaning that even if the most promising surrounding parameters do not meet their prediction expectations, other surrounding parameters that are used as inputs in this research could also be easily applied. Two good example of surrounding parameters which are likely to be much more significant for DH networks than they were for water and gas are the road coverage (percentage asphalt) and the average distance from a house in the project area to the street. The road coverage is likely to be more significant because a bigger part of the network will be below the street (instead of sidewalks). The average distance to the street is more promising because in the water and gas replacement projects most of the time the connection pipes were not replaced. However, for a new DH network connection pipes are required and likely to be more expensive if houses are further away from the network.

As for the third contribution at this point, it is hard to say whether using the gathered gas and water data as extra training data for a heat model is going to be of added value. This will have to be validated when heat data is available. Based on the experience in this research, it is a 50/50 chance, since using extra water data for gas predictions improved the modeling performance significantly, but using extra gas data for water prediction did not.

Does including surrounding parameters in a statistical model, calculating construction costs of pipe infrastructures in the built environment, improve the model performance, and can a trained model and or project data from similar infrastructures be used to increase the model performance of a district heating model?

Summarizing the answers of the subquestions the following main conclusion can be drawn. Current energy transition models are not able to accurately calculate construction costs of DH networks because they do not (sufficiently) consider surrounding parameters. Linear regression is the most promising ML approach to include surrounding parameters in a model that calculates construction costs per meter of (fluid) infrastructures in the built environment. Based on interviews with experts 25 potential surrounding parameters that are likely to influence the construction costs are identified. When applying linear regression it was found that the pollution category (surrounding parameter) and 1/(total length) (no surrounding parameter) are the two most promising input parameters besides the average diameter for predicting construction costs of water and gas replacement projects. Including these, and other input parameters, resulted in significantly better modeling performances which led to a bandwidth reduction of **25%** with respect to the chosen baseline (only diameter) for both the water and gas model. Also the MSE for both models decreased significantly (34% gas and 35% water) indicating that the significant increase in model performance is not dependent on the chosen main scoring criteria (R^2). The results of this research are promising for DH cost predictions in three following ways: to act as a proof of concept for the modeling approach, to identify the most important surrounding parameters, and to provide potential extra training data. To get the most out of this analysis for DH it is important to keep in mind the five most important differences stated below subquestion 4.

All in all, it is believed that despite the differences between the three considered infrastructures, using the methodology, and potentially the data, provided by this research will be a valuable contribution to energy transition models by generating more accurate cost key figures. Which in turn is a valuable contribution to the dutch energy transition in the built environment since these energy transition models are being used by municipalities that are responsible for shaping this transition.

8.2. Recommendations

In this final section of the report, some recommendations are given. The recommendation section is divided into two parts. First, in subsection 8.2.1 potential improvements for the modeling strategy applied in this specific research are presented. Secondly, broader recommendations on potential future research related to this research are discussed in subsection 8.2.2.

8.2.1. Potential model improvements

In this subsection, six potential improvements to the model developed in this research are proposed. The first proposition is applying different kinds of ML methods instead of linear regression, starting with the same total input table containing all the potential inputs. A potential advantage of this approach is that certain ML algorithms automatically select the best inputs. An example of such a model that might be a good alternative to linear regression is the decision tree variant random forests. A random forest approach also applies bagging and the results from this research (both best performing models apply bagging) suggest this is a promising method to predict construction costs.

The second proposition is to alter the data prepossessing process described in section 5.1. Examples of possible alterations are choosing different categories from the house type, neighborhood type, or road type. But also several inputs can be combined to create categorical inputs like historical city centers or shopping areas. Finally, the combination of the material type and diameter can be entered in a different way than the currently used 9 categories that are described in Table 5.2.

A third proposition is to implement interaction terms in the ML model. Based on the literature review, adding interaction terms is a promising addition to linear regression. However, because of the many potential inputs and therefore even more potential combinations of these inputs in this research, there was, unfortunately, no time left to implement this.

Furthermore, a fourth suggestion is to alter the applied bagging algorithm. Instead of using the average of all the m models used for bagging, a weighted average based on the modeling performance can be used. The idea behind using this weighted average is that bad performing models have a smaller influence and will therefore be less likely to decrease the total model performance. It is important to note that the model performance is the $R2\_test$ score and not the $R2\_train$ score. In the current modeling setup, two potential data set candidates can be used as the test data set. The first candidate is the main test set that originates from the k-fold splitting process and is used to score the performance of the resulting (total) model. When using this data set as the test data set used in the weighted average the model is more prone to overfitting. Since the entire purpose of using a test and train dataset is to reduce the risk of overfitting it is not recommended to use this main test data set. The second candidate for a test data set used for the weighted average is the data set containing the projects that are not used in a specific bagging model. Every bagging model is trained using the same amount of data points as the main test set resulting from k-fold splitting. However when creating the training set for a specific bagging model certain data points from the main test set can be used multiple times. This means that other data points from the main training set are not used for this specific bagging model. Using all these "unused" data points to score the resulting bagging models is the recommended solution for taking the weighted average.

A fifth suggestion is, to try to remove the parameters that are added by the algorithm to the ML model based on random chance. Two suggestions for realizing this are: First, find a way to optimize the model without using a random seed for splitting the total dataset, to reduce the influence of the splitting of the test and training set, to the optimized model performance. A complicating factor is that the model performance could decrease or increase not because the modeling settings or inputs were changed but because the splitting of the test and train set was altered. Second, use the t-stat of the chosen input parameters in the decision-making process of the algorithm. Instead of only looking at the realized model performances.

The final proposition to improve the developed model is to further investigate the reason why data providing companies are important input parameters for predicting construction costs. In the ideal situation, these input parameters are not present in the best performing models. Removing these input parameters from the final models is desirable because the generated cost key figures are not applicable in high-level energy transition models when the responsible company is a required input. A suggestion for this would be to remove the data providing companies as potential inputs for the algorithm and to see what happens to the resulting models. When the modeling performance of the final models only decreases by a small amount, the models without the data providing companies as model inputs, are probably better-suited models even though their performance is a little less.

8.2.2. What should happen next?

In this subsection, three recommendations are given for promising future initiatives that are related to the research conducted in this thesis. The first recommendation, which is rather general, is that more awareness for the lack of available cost data for DH networks in the Netherlands should be created. During this research, it was realized that, especially in scientific literature, there is very little information available about construction costs of district heating networks in general and even less about construction costs in the Netherlands. The fact that there is so little scientific literature is probably caused by the lack of available (open source) data that could be used as input for potential studies. The large scale implementation of DH networks is a serious candidate for the natural gas transition in built environment. However, the potential is of DH networks is undermined because of the lack of cost data. More awareness of the lack of detailed cost data for district heating networks is very important to solve this problem and increase the district heating potential. A lack of detailed cost data is a problem because choosing locations for new district heating networks is a lot more challenging and therefore less likely to happen on a big scale.

Additionally, a second suggestion is to further validate how the results of this analysis can be used to model district heating networks. Things that could be looked into are:

- Are the same surrounding parameters indeed significant for cost predictions?
- Look into which surrounding parameters have similar influences on the construction costs for different infrastructures. Use the data from water and gas to specifically train the parameters which seem to have a similar influence on the construction cost.
- Further look into the concept of transfer learning. *"Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned"* [58]. Apply this concept to improve the learning of a ML model predicting construction costs of DH networks using the knowledge from the developed water and gas models. It should be noted that transfer learning is particularly useful for sequential model structures like for example neural networks [38]. Meaning that it is probably desirable to choose a different ML approach than linear regression when transfer learning is applied.

- Validate that using the same modeling approach as applied in this research also significantly reduces the bandwidth for construction costs predictions of DH networks when using heat data to train the model.

A final recommendation is to look into the potential of sewer systems for the cost prediction of DH networks. Even though sewer systems are not pressurized pipe infrastructures like water, gas, and heat, the average diameter and location in the street profile (below the street) of sewer systems are reasons to believe that sewer systems in fact are very similar to DH networks. At least for these two aspects, sewer systems are a lot more similar than water and gas infrastructures and when using specific infrastructures to train specific model parameters, as discussed in the previous recommendation, sewer systems are a promising candidate to include in the analysis.

Bibliography

[1] Advanced: Heat infrastructure costs. https://docs.energytransitionmodel.com/main/heat-infrastructure-costs. Accessed: 2020-05-28.

[2] hat is the share of renewable energy in the eu? https://ec.europa.eu/eurostat/cache/infographs/energy/bloc-4c.html. Accessed: 2020-09-10.

[3] Reference. https://geopandas.org/reference.html. Accessed: 2020-12-17.

[4] Flexalen 1000+. https://thermaflex.com/en/products/pre-insulated-pipes/flexalen-1000, . Accessed: 2020-10-01.

[5] Thermos help page. https://tool.thermos-project.eu/help/index.html, . Accessed: 2020-07-02.

[6] Transitievisie warmte en wijkuitvoeringsplan. https://www.rvo.nl/onderwerpen/duurzaam-ondernemen/duurzame-energie-opwekken/aardgasvrij/aan-de-slag-met-aardgasvrij/transitievisie-warmte-en-wijkuitvoeringsplan. Accessed: 2020-07-30.

[7] Wat mag ik vragen voor het leveren van warmte? https://www.acm.nl/nl/warmtetarieven. Accessed: 2021-01-27.

[8] Clip. https://desktop.arcgis.com/en/arcmap/10.3/tools/analysis-toolbox/clip.htm. Accessed: 2020-12-16.

[9] Greenhouse gases. https://longreads.cbs.nl/european-scale-2019/greenhouse-gases/. Accessed: 2020-05-01.

[10] statsmodels.regression.linear_model.ols.fit_regularized. https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.fit_regularized.html. Accessed: 2021-01-03.

[11] Kabinet: einde aan gaswinning groningen. https://www.rijksoverheid.nl/actueel/nieuws/2018/03/29/kabinet-einde-aan-gaswinning-in-groningen. Accessed: 2020-05-01.

[12] Waar komen de warmtenetten? https://www.natuurenmilieu.nl/themas/energie/projecten-energie/nieuwe-warmte/warmtenetten/warmtenetten-waar/?gclid=CjwKCAiAt9z-BRBCEiwA_bWv-CNLUlZ52hwNJ54fJotBjHPcsxzID5hJsUeIFMVRQpNH7VNUfKBQZRoCR1IQAvD_BwE. Accessed: 2020-12-14.

[13] 2019 gas. https://www.netbeheernederland.nl/_contentediting/files/files/EN_Gas-2019-Legenda.pdf, . Accessed: 2020-12-15.

[14] Waterleveranciers overzicht. https://www.easyswitch.nl/waterleveranciers-overzicht/, . Accessed: 2020-12-15.

[15] Paris agreement. https://ec.europa.eu/clima/policies/international/negotiations/paris. Accessed: 2020-05-01.

[16] Linear regression. https://www.statsmodels.org/stable/regression.html#regression--page-root. Accessed: 2021-01-03.

[17] Energy report transition to sustainable energy. Technical report, Ministry of economic affairs, The hague, 2016. URL https://www.government.nl/documents/reports/2016/04/28/energy-report-transition-tot-sustainable-energy.

[18] Vertrouwen in de toekomst Regeerakkoord 2017-2021. Technical report, Rijksoverheid, The hague, 2017. URL https://www.rijksoverheid.nl/regering/documenten/publicaties/2017/10/10/regeerakkoord-2017-vertrouwen-in-de-toekomst.

[19] Helge Averfalk and Sven Werner. Economic benefits of fourth generation district heating. *Energy*, 193: 116727, 2020. ISSN 03605442. doi: 10.1016/j.energy.2019.116727. URL https://doi.org/10.1016/j.energy.2019.116727.

[20] Robert E. Best, P. Rezazadeh Kalehbasti, and Michael D. Lepech. A novel approach to district heating and cooling network design based on life cycle cost optimization. *Energy*, 194:116837, 2020. ISSN 03605442. doi: 10.1016/j.energy.2019.116837. URL https://doi.org/10.1016/j.energy.2019.116837.

[21] Subhes C. Bhattacharyya and Govinda R. Timilsina. A review of energy system models. *International Journal of Energy Sector Management*, 4(4):494–518, 2010. ISSN 17506220. doi: 10.1108/17506221011092742.

[22] Benno Boeters. Warmtenetten in tien jaar drie keer zo groot. *Technischweekblad.nl*, 2019. Accessed: 2020-05-01.

[23] Simone Buffa, Marco Cozzini, Matteo D'Antoni, Marco Baratieri, and Roberto Fedrizzi. 5th generation district heating and cooling systems: A review of existing cases in Europe. *Renewable and Sustainable Energy Reviews*, 104(December 2018):504–522, 2019. ISSN 18790690. doi: 10.1016/j.rser.2018.12.059. URL https://doi.org/10.1016/j.rser.2018.12.059.

[24] Ronson Chee, Kevin Lansey, and Erickson Chee. Estimation of water pipe installation construction costs. *Journal of Pipeline Systems Engineering and Practice*, 9(3):1–16, 2018. ISSN 19491204. doi: 10.1061/(ASCE)PS.1949-1204.0000323.

[25] Robert M. Clark, Mano Sivaganesan, Ari Selvakumar, and Virendra Sethi. Cost Models for Water Supply Distribution Systems. *Journal of Water Resources Planning and Management*, 128(5):312–321, 2002. ISSN 0733-9496. doi: 10.1061/(asce)0733-9496(2002)128:5(312).

[26] D. Connolly, H. Lund, B. V. Mathiesen, and M. Leahy. A review of computer tools for analysing the integration of renewable energy into various energy systems. *Applied Energy*, 87(4):1059–1082, 2010. ISSN 03062619. doi: 10.1016/j.apenergy.2009.09.026.

[27] David Connolly, Brian Vad Mathiesen, Poul Alberg Ostergraad, Bernd Möller, Steffen Nielsen, Henrik Lund, Urban Persson, and Sven Werner. Heat Roadmap Europe 2. Technical report, Aalborg University, 2013. URL https://vbn.aau.dk/ws/files/77342092/Heat_Roadmap_Europe_Pre_Study_II_May_2013.pdf.

[28] A. Dalla Rosa, R. Boulter, K. Church, and S. Svendsen. District heating (DH) network design and operation toward a system-wide methodology for optimizing renewable energy solutions (SMORES) in Canada: A case study. *Energy*, 45(1):960–974, 2012. ISSN 03605442. doi: 10.1016/j.energy.2012.06.062. URL http://dx.doi.org/10.1016/j.energy.2012.06.062.

[29] Jasper Donker and Tanneke Ouboter. SRP Energy Transition – DiDo comparison with other energy models. (January), 2015.

[30] EBN. Energy in numbers 2020, 2020. URL https://www.energieinnederland.nl/wp-content/uploads/2020/02/EBN-INFOGRAPHIC-2020-ENG.pdf.

[31] Tobias Fleiter, Rainer Elsland, Matthias Rehfeldt, Jan Steinbach, Ulrich Reiter, Giacomo Catenazzi, Martin Jakob, Cathelijne Rutten, Robert Harmsen, Florian Dittmann, Philippe Riviere, and Pascal Stabat. EU Profile of heating and cooling demand in 2015. *HeatRoadmapEU*, (695989):70, 2017. URL http://heatroadmap.eu/output.php.

[32] Daniel Fredrik Hedenus, Fredrik Hedenus, Daniel Johansson, and Kristian Lindgren. A Critical Assessment of Energy - economy - climate Models for Policy Analysis. *Journal of Applied Economics and Business Research*, 3(2):118–132, 2013. ISSN 1927-033X.

[33] Dr. Ir. P. W. Heijnen. Statistical modelling Reader EPA1314. Technical Report 1, Delft university of Technology, Delft, 2015.

[34] Piotr Hirsch, Kazimierz Duzinkiewicz, Michał Grochowski, and Robert Piotrowski. Two-phase optimizing approach to design assessments of long distance heat transportation for CHP systems. *Applied Energy*, 182:164–176, 2016. ISSN 03062619. doi: 10.1016/j.apenergy.2016.08.107. URL `http://dx.doi.org/10.1016/j.apenergy.2016.08.107`.

[35] Nico Hoogervorst, Tijs Langeveld, Bas van Bemmel, Folckert van der Molen, Steven van Polen, and Ruud van den Wijngaart. Startanalyse aardgasvrije buurten. Technical report, PBL, Den Haag, 2019. URL `https://www.pbl.nl/publicaties/startanalyse-aardgasvrije-buurten-2020`.

[36] Ch Hueber, K. Horejsi, and R. Schledjewski. Review of cost estimation: methods and models for aerospace composite manufacturing. *Advanced Manufacturing: Polymer and Composites Science*, 2(1): 1–13, 2016. ISSN 20550359. doi: 10.1080/20550340.2016.1154642. URL `http://dx.doi.org/10.1080/20550340.2016.1154642`.

[37] Ed Kerckhoffs. Werking Caldomus Werking Caldomus. Technical report, Innoforte, Druten, 2017. URL `https://www.innoforte.nl/in/wp-content/uploads/2018/02/171209-werking-Caldomus.pdf`.

[38] Tanya Kolosova and Samuel Berestizhevsky. Supervised Machine Learning. *Supervised Machine Learning*, 2020. doi: 10.1201/9780429297595.

[39] Martin Leurent, Pascal Da Costa, Miika Rämä, Urban Persson, and Frédéric Jasserand. Cost-benefit analysis of district heating systems using heat from nuclear plants in seven European countries. *Energy*, 149:454–472, 2018. ISSN 03605442. doi: 10.1016/j.energy.2018.01.149.

[40] Wen Liu, Diederik Klip, William Zappa, Sytse Jelles, Gert Jan Kramer, and Machteld van den Broek. The marginal-cost pricing for a competitive wholesale district heating market: A case study in the Netherlands. *Energy*, 189, 2019. ISSN 03605442. doi: 10.1016/j.energy.2019.116367.

[41] Francisco Lobato, Frank Coumans, Huub Heygele, Debora Kuipers, and Jan Rours. VeWa Veiligheidsvoorschrift Warmte. Technical report, Energie-Nederland, Arnhem, 2015. URL `https://www.energie-nederland.nl/app/uploads/2017/02/150921VeWa2015-Definitief.pdf`.

[42] Jean Loup Loyer, Elsa Henriques, Mihail Fontul, and Steve Wiseall. Comparison of Machine Learning methods applied to the estimation of manufacturing cost of jet engine components. *International Journal of Production Economics*, 178:109–119, 2016. ISSN 09255273. doi: 10.1016/j.ijpe.2016.05.006. URL `http://dx.doi.org/10.1016/j.ijpe.2016.05.006`.

[43] Henrik Lund, Sven Werner, Robin Wiltshire, Svend Svendsen, Jan Eric Thorsen, Frede Hvelplund, and Brian Vad Mathiesen. 4th Generation District Heating (4GDH). Integrating smart thermal grids into future sustainable energy systems., 2014. ISSN 03605442.

[44] Rasmus Lund and Soma Mohammadi. Choice of insulation standard for pipe networks in 4th generation district heating systems. *Applied Thermal Engineering*, 98:256–264, 2016. ISSN 13594311. doi: 10.1016/j.applthermaleng.2015.12.015. URL `http://dx.doi.org/10.1016/j.applthermaleng.2015.12.015`.

[45] Valentina Marchionni, Nuno Lopes, Luis Mamouros, and Dídia Covas. Modelling Sewer Systems Costs with Multiple Linear Regression. *Water Resources Management*, 28(13):4415–4431, 2014. ISSN 09204741. doi: 10.1007/s11269-014-0759-z.

[46] Valentina Marchionni, Marta Cabral, Conceição Amado, and Dídia Covas. Estimating water supply infrastructure cost using regression techniques. *Journal of Water Resources Planning and Management*, 142(4), 2016. ISSN 07339496. doi: 10.1061/(ASCE)WR.1943-5452.0000627.

[47] Netbeheer Nederland. rekenmodelen overzicht. Technical report, The hague. URL `https://www.netbeheernederland.nl/_upload/Files/Rekenmodellen_21_836d4f1302.pdf`.

[48] Overmorgen. Het Warmtetransitiemodel. Technical report, Overmorgen, Amesfoort, 2020.

[49] Yu Pan, Liuchen Liu, Tong Zhu, Tao Zhang, and Junying Zhang. Feasibility analysis on distributed energy system of Chongming County based on RETScreen software. *Energy*, 130:298–306, 2017. ISSN 03605442. doi: 10.1016/j.energy.2017.04.082. URL `http://dx.doi.org/10.1016/j.energy.2017.04.082`.

[50] Urban Persson, Eva Wiechers, Bernd Möller, and Sven Werner. Heat Roadmap Europe: Heat distribution costs. *Energy*, 176:604–622, 2019. ISSN 03605442. doi: 10.1016/j.energy.2019.03.189. URL `https://www.sciencedirect.com/science/article/pii/S0360544219306097`.

[51] Platform Geothermie, ebn, Dutch association geothermie operators, and Stichting warmtenetwerk. Masterplan Aardwarmte in Nederland. Technical report, 2018. URL `https://www.ebn.nl/wp-content/uploads/2018/05/20180529-Masterplan-Aardwarmte-in-Nederland.pdf`.

[52] C Reidhav and Sven Werner. Investment models for district heating in areas with detached houses. *10th International Symposium on District Heating and Cooling*, (September), 2006.

[53] RHDHV. CoP Kostencalculatie. Technical report, RHDHV Water Technology B.V., Amesfoort, 2014.

[54] Benno Schepers, Ruud van den Wijngaart, Alexander Oei, and Maarten Hilferink. Functioneel ontwerp Vesta 4.0. Technical report, CE Delft, Delft, 2019. URL `https://www.pbl.nl/sites/default/files/downloads/pbl-2019-ce-delft-functioneel-ontwerp-vesta-4.0_4085.pdf`.

[55] Tariq Shehab, Elhami Nasr, and Mohammad Farooq. Conceptual cost estimating model for water and sewer projects. *Pipelines 2014: From Underground to the Forefront of Innovation and Sustainability - Proceedings of the Pipelines 2014 Conference*, (Asce 2013):367–373, 2014. doi: 10.1061/9780784413692.033.

[56] Rifat Sonmez. Conceptual cost estimation of building projects with regression analysis and neural networks. *Canadian Journal of Civil Engineering*, 31(4):677–683, 2004. ISSN 03151468. doi: 10.1139/L04-029.

[57] Takuya Togawa, Tsuyoshi Fujita, Liang Dong, Minoru Fujii, and Makoto Ooba. Feasibility assessment of the use of power plant-sourced waste heat for plant factory heating considering spatial configuration. *Journal of Cleaner Production*, 81:60–69, 2014. ISSN 09596526. doi: 10.1016/j.jclepro.2014.06.010. URL `http://dx.doi.org/10.1016/j.jclepro.2014.06.010`.

[58] Lisa Torrey and Jude Shavlik. Transfer Learning. Technical report, University of Wisconsin, Wisconsin, 2010.

[59] Harm Valk, T Haytink, J Kaspers, P van Meegeren, and J Zijlstra. Verkenning tool aardgasvrije bestaande woningen In opdracht van het ministerie van Economische Zaken en Klimaat. Technical report, Nieman, Utrecht, 2018. URL `https://www.rvo.nl/sites/default/files/2018/06/Rapportverkennendestudietoolaardgasvrijewoningen_0.pdf`.

[60] Jun Wang and Baabak Ashuri. Predicting ENR Construction Cost Index Using Machine-Learning Algorithms. *International Journal of Construction Education and Research*, 13(1):47–63, 2017. ISSN 15503984. doi: 10.1080/15578771.2016.1235063. URL `http://dx.doi.org/10.1080/15578771.2016.1235063`.

[61] Warming UP. Definities warmtebranche. Technical report, 2020.

[62] WarmingUP. WarmingUP Innovatieplan – samenvatting Deelnemers en derden. Technical Report december, WarmingUP, Delft, 2019. URL `https://warmtenetwerk.nl/wp-content/uploads/Samenvatting-Innovatieplan-WarmingUP.pdf`.

[63] Larry Wasserman. *All of Nonparametric Statistics*, volume 102. Springer Texts in Statistics, New York, 2006.

[64] Sven Werner. International review of district heating and cooling. *Energy*, 137:617–631, 2017. ISSN 0360-5442. doi: 10.1016/j.energy.2017.04.045. URL `https://doi.org/10.1016/j.energy.2017.04.045`.

A

Machine learning principles explained

A.0.1. k-fold cross validation

To be able to calculate the prediction accuracy of a ML model the data set needs to be divided into a test (y_{test},x_{test}) and train (y_{train},x_{train}) data set. For smaller data sets the performance of the model can vary quite significantly depending on which part of the data ends up in which subset as shown in Table 5.4. To get a better feeling of the real (average) performance of a certain modeling approach it is common to divide the data set into the two subsets multiple times. A common way of doing this is by k-fold cross validation also known as, k-fold splitting. K-fold splitting randomly splits the total database in k different data sets containing all the input parameters but only ($\frac{n}{k}$) projects. Every generated database is used ones for validation and $(k-1)$ times for training. This means that the higher the number k the smaller the testing data set and visa versa. In the most extreme scenario, choosing $k = 2$ the test and train data sets are the same size. In the other extreme scenario where $k = n$ the test data set only contains a single project.

As mentioned k-fold splitting randomly divides all the data points over the k different subsets. The fact that this splitting is random is very important because it will make sure that potential data shorting of the input table does not have any effect on the splitting of the database. For example, if all the projects in the input table would be sorted based on when they were realized, and the splitting was not random then all the projects used for training might be realized in 2018 and all the projects used for testing are realized in 2019. If the moment of realization would have an impact on the construction cost that would be an undesirable situation. To make sure that any patterns in the input data, known and unknown to the data engineer, do not have an effect on the scoring random splitting is important. Another example of a pattern that should be checked when splitting the data for this specific research is the company that realized the project. It is quite likely that certain companies, in general, have lower or higher costs than other companies, which is not necessarily caused by the surrounding. Because of this reason, it is important that all companies have some data in the test set and some data in the training set. Since not all the companies have provided the same amount of projects it could be that when randomly splitting the data a company that provided fewer projects is only present in one of the two subsets.

It could also be that it is actually desired to only have certain specific projects in your test data set. In this project that is the case when the combined models are trained and scored. When training the model data from both water and gas should be used. But for the combined water model, the test dataset, should only contain water projects since your not interested in the construction cost of gas but your only using the gas data to increase the sample size. In this case, it is still important to randomly select the projects. However, the random selection is only made from the available water projects. Originally this was a very important aspect for the final combined model that was supposed to predict heat construction cost based on water, gas, and heat data. However, since no heat data is available it is not possible to create a test data set containing heat data. To check if using data from a similar infrastructure could be of added value the combined models for water and gas are developed. If the combined models outperform the normal models this means that it is also more promising to use water and gas data for DH cost predictions.

A.0.2. Bagging

The idea of bagging is relatively simple. Instead of training a single model and using this model to predict the dependent variable, when bagging is applied $m_{bagging}$ different models are trained and the average of all the predictions of the $m_{bagging}$ models is used as the final prediction. The advantage of bagging is that it can reduce the variance of a model without increasing the bias. Generally speaking, models that tend to overfit data have high variance and are therefore very suited for bagging. Bagging in this case compensates a bit for the overfitting of training data because the different models are trained on different training datasets. Because of this reason in order for bagging to be successful, it is very important that the training dataset of all the $m_{bagging}$ different models are randomly chosen. Meaning that it could be that some data points are used to train every single model whereas other data points are never used for training. This random selecting of the training data, where the data engineer is also able to choose the size of the training dataset, is what distinguishes Bagging from k-fold splitting. When applying k-fold splitting every data point is used once for validation and (k-1) times for training. Also for k-fold splitting, the size of the training dataset is dependent on the number of splits and can not be chosen freely by the data engineer. A ML model that uses bagging can be scored in two different ways. The first option is to score all the $m_{bagging}$ separate models using the data that is not used for training a specific model. Taking the average of all the R^2 scores of the different models gives an indication of the lower bound of the performance of the total model. The upside of this is that all the available data is used to train the model and still a reasonable prediction can be made about the model performance. The second option is to first apply k-fold splitting and then apply the bagging to the k different datasets. The total model prediction resulting from the $m_{bagging}$ different models used in the bagging approach can then be validated with the test dataset containing data that is not used for any of the $m_{bagging}$ models. The advantage of this second approach is that you have a more accurate prediction of the model performance. However since less data is used to train the model, the model is likely to perform less compared to the first approach where all data is used.

A.0.3. Tolerance value

When identifying potential new inputs for the model one could look at the correlation of all potential inputs with the dependent variable. Besides looking at the correlation of potential input parameters with the dependent variable (y) it is also important to get a feeling for the amount of correlation that the proposed input parameter has with the input parameters that are already in the model. A good measure of this is the tolerance value. When the tolerance value of a potential input parameter is high this means that the variation in the new input parameter can not be explained by the parameters already in the model. In other words, this means that the new input parameter adds a lot of new information to the model and is therefore a good new input. The tolerance value of a potential input parameter can be calculated by considering this input parameter as the dependent variable of a regression model containing the input parameters that are already in the model. For this model the R^2 needs to be calculated, subtracting this R^2 value from 1 results in the tolerance value ($TOL = 1 - R^2$)[33].

A.0.4. Regularization

A common method to reduce the number of input parameters in machine learning is to apply so-called regularization. When using regularization the cost function of the ML model is altered. The cost function is the function that is minimized during the training process of the model. The cost function is the average of the loss function results. The loss function is calculated for every training data point. The loss function can have many different forms and is dependent on the type of ML that is applied. For linear regression, the most common loss function, also applied in this research, is the squared error loss shown in equation Equation 2.9.

There are different types of regularization which "regulate" a ML model in a certain way. The common regularization methods that are considered in this research are L^2 regularization, L^1 regularization, and a combination of both of these. With these regularization methods, a penalty term is added to the loss function. This penalty term is a certain factor (λ) multiplied with the regression parameters ($\boldsymbol{\theta}$). This regularization factor λ is a user input. The higher this value the stronger the regularization. When this factor is chosen too high all the regression parameters will become zero and the model will not work anymore. When the factor is chosen to low, the regularization has limited effect, in the extreme situation where $\lambda = 0$ the regularization has no effect at all and a "normal" linear regression model is trained. The difference between L^1 and L^2 regularization is the exponent of θ in the penalty term. When using L^2 regularization the square of θ is used as a penalty term and when using L^1 the regular θ is used. The loss functions of normal linear regression, linear regression using L^1 regularization and regression using L^2 regularization are presented in Equation A.1, Equation A.2 and Equation A.3 respectively [38]. Where $\hat{\boldsymbol{\theta}}$ represents the estimated regression parameters, the goal of training a linear regression model.

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\min} \frac{1}{n} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 \tag{A.1}$$

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\min} \frac{1}{n} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \tag{A.2}$$

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\min} \frac{1}{n} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \tag{A.3}$$

The effects of the two different regularization terms are different. The L^2 tends to decrease the value of all the regression parameters, which can lead to a more robust model. The L^1 regularization tends to make the regression parameters (θ_i) of less important inputs (x_i) equal to zero. Because of this phenomenon, L_1 regularization is a useful tool when selecting the most important inputs. It is also possible to use a combination of both regularization methods by adding both penalties terms together, this is also known as elastic net regularization. When combining both regularization methods and extra regularization term $L1_{wt}$ is included. $L1_{wt}$ regulates the proportion of the two different regularization types. The loss function of elastic net regularization is shown in Equation A.4 [10]. It can be seen that when $L1_{wt}$ is set to 1 elastic net regularization turns into L_1 regularization and when $L1_{wt}$ is set to 0 elastic net regularization turns into L_2 regularization. For that reason, only the elastic net regularization function with the two user inputs $L1_{wt}$ and λ is used for the model developed in this research.

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\min} \frac{1}{n} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda \left(\frac{1 - L1_{wt}}{2} |\boldsymbol{\theta}\|_2^2 + L1_{wt} |\boldsymbol{\theta}\|_1 \right) \tag{A.4}$$

A.0.5. Interaction terms

A possible adaption to input data are so-called interaction terms. As already identified in subsection 2.2.1 interaction terms are quite common in linear regression models used for similar infrastructures. An interaction term is created by multiplying two or more inputs with each-other to generate a single new input. When an interaction term is used it is important for the interaction term to work properly that all the inputs that are multiplied to create the interaction term are also inputted separately into the model. An interaction term should be used when it is suspected that the relation of an independent variable (x_1) with the dependent variable (y) is dependent on another independent variable (x_1). An example of this would be that the relation between the amount of meters and the construction cost is dependent on the average diameter. It could be that for smaller diameters the total amount of meters is less important than for bigger diameters or the other way around.

A.0.6. Non-linear input parameter transformations

A possible input data alteration is to use non-linear transformations of your original input data to make the linear regression model non-linear. This was already described in subsection 2.2.1 and an example was given in Equation 2.10 and Equation 2.11. A few examples of commonly used non-linear transformations are: $Log(x), \sqrt{x}, \frac{1}{x}, x^i$. These non-linear transformations should be used when it is expected that a certain input parameter (x_i) has a specific non-linear relation with the dependent variable (y). This expectation can arise when residual plots are studied. Also, it is possible to identify a certain non-linear relation by looking at a simple scatter plot of a certain input parameter. Another reason why data engineers sometimes use a $Ln(x)$ or $Ln(y)$ transformation is that the parameters should be normally distributed. Using an $Ln()$ transformation can make a dataset that wasn't normally distributed before a normal distribution. The downside from using a $Ln()$ transformation, especially when doing it to the dependent variable (y), is that the interpretability of the models decreases.

B
Conducted interviews

A big part of this research is based on conducting interviews. In total interviews with 50 different people from 34 different companies are conducted. With some people, mostly from data providing companies, two different interviews are conducted. In the first interview questions, related to the different research questions, were asked and the cooperation opportunities were identified and the second interview was used to specifically talk about data gathering (spreadsheet). In the Table B.1 an overview of all the conducted interviews is presented. In the table, the name, company, and expertise of the respondent are given together with the date on which the interview took place.

[1] Questions after a webinar instead of interview

Table B.1: Overview of conducted interviews

Naam	Bedrijf	Topic (expertise)	Datum	Datum 2
Peter de grave	Deltares	Build database for model input	26-3-2020	
Lefki Loverdou	Deltares	District heating model GIS	02-4-2020	
Alex Koster	Deltares	Cost calculations (dikes)	06-4-2020	
Max brouwer	Zuid holland province	Energy transition models	20-4-2020	
Arne Bosch	Waternet	Drinking water	30-4-2020	18-8-2020
Jan Peter van der Hoek	TU Delft	Drinking water	30-4-2020	
Ryvo Octaviano	TNO	Energy transition models	06-5-2020	
Geert Linsen	WML	Drinking water	08-5-2020	01-9-2020
Roel Diemel	Brabantwater	Drinking water	11-5-2020	03-9-2020
Kurt Marlein	Comsof	Energy transition models[1]	12-5-2020	
Peter Horst	PWN	Drinking water	14-5-2020	11-9-2020
Peter van Houwelingen	Oasen	Drinking water	15-5-2020	21-8-2020
Rik Verweij	Stedin	Natural gas / district heating	18-5-2020	
Rob Cloosen	Stedin	Natural gas / district heating	18-5-2020	
Bobby Ham	Platform Dio	Drinking water tendering	18-5-2020	
Lennard Wools	Evides	Drinking water	22-5-2020	02-9-2020
Eric Can	Dunea	Drinking water	25-5-2020	21-8-2020
Marcel Bakker	RHDHV	Kostenstandaard drinkwater	29-5-2020	
Berend Doedens	Heijmans	Contracter	03-6-2020	
Ronald Roosjen	Deltares	GIS data	09-6-2020	
Gerrit Hendriksen	Deltares	GIS data	12-6-2020	
Ivo Smits	Qirion	Energy transition models	16-6-2020	
Pepijn Caers	Heijmans	Contracter (cost engineer)	16-6-2020	
Linda Maring	Deltares	Spatial planning (subsurface)	23-6-2020	
Matthijs Hansen	Witteveen en Bos (thesis)	Energy transition models	22-7-2020	
Sjoerd Braaksma	Gemeente Rotterdam	Tools for energy transition	29-7-2020	
Rens Reiff	Darel	Energy transition models	06-8-2020	
Gerard van Roo	Liander	Natural gas	10-8-2020	
Menno Karres	Waternet	Drinking water	17-8-2020	
Jasper Selten	Goconnectit	Drinking water / Natural Gas	17-8-2020	
Corina van der Hulst	Stedin	Natural gas	18-8-2020	
Roald Leemrijse	WMD	Drinking water	20-8-2020	
Sjoerd Buddingh	Alliander	Natural gas	21-8-2020	
Henk Tijssens	Waternet	Drinking water	25-8-2020	
Berend Nootenboom	Stedin	Natural gas / district heating	25-8-2020	
Theo Ellenbroek	COB	Spatial planning (subsurface)	31-8-2020	
Steffen Nielsen	Aalborg Uni (thermos)	Energy transition models	09-9-2020	
Rudi Zoet	Goconnectit	Cost modeling with surrounding	14-9-2020	
Aldo Veneman	Vitens	Drinking water	21-9-2020	
Bernard Enthoven	Waterbedrijf Groningen	Drinking water	23-9-2020	
Bob Goessen	Stedin	Natural gas	01-10-2020	06-11-2020
Theo Venema	Warmtestad Groningen	District heating / drinking water	05-10-2020	
Tugay Akbulut	ACM	District heating cost gathering	12-10-2020	
Ko Spruit	Waternet	Drinking water	13-10-2020	17-11-2020
Kathelijne Bouw	Hanzehogeschool Phd	Energy transition models	21-10-2020	
Henk Tuinstra	Waterbedrijf Groningen	Drinking water	27-10-2020	
Petra Heijnen	TU delft	Machine learning	29-10-2020	
Arjan Hekker	Enexis	Natural gas	12-11-2020	20-11-2020
Luis Sanchez Garcia	Halmstad university (Phd)	District heating	02-12-2020	
Marina Vasarini Lopes	Gemeente Amsterdam	Spatial planning (subsurface)	04-12-2020	

C

Extra results Gas

Table C.1: Model scores gas model, tyring different lambda's for L1 regularization

Model settings: n_kfold=5, Regularization=True, reg_L1_wt = 1, Bagging=False

Inputs Model:	Total length, pollution, inhabitants, diameter, network opr 1			Total length, pollution, inhabitants, diameter, network opr1, footprint building			Total length, pollution, inhabitants, diameter, network opr 1, footprint building, PVC_medium			Total length, pollution, inhabitants, diameter, network opr 1, footprint building, PVC_medium, Percentage trees		
Lambda	**1**	**3**	**6**	**1**	**3**	**6**	**1**	**3**	**6**	**1**	**3**	**6**
train avg	0.59	0.58	0.57	0.59	0.58	0.57	0.61	0.60	0.58	0.62	0.61	0.60
test min	-0.09	-0.08	-0.08	-0.07	-0.08	-0.08	-0.31	-0.34	-0.35	-0.30	-0.32	-0.35
test max	0.68	0.66	0.63	0.68	0.66	0.63	0.62	0.60	0.61	0.59	0.59	0.57
Test AVG	**0.37**	**0.37**	**0.36**	**0.37**	**0.37**	**0.36**	**0.28**	**0.28**	**0.27**	**0.23**	**0.24**	**0.23**

Table C.2: Model scores gas model, tyring different lambda's for L2 regularization

Model settings: n_kfold=5, Regularization=True, reg_L1_wt = 0, Bagging=False

Inputs Model:	Total length, pollution, inhabitants, diameter, network opr 1			Total length, pollution, inhabitants, diameter, network opr1, footprint building			Total length, pollution, inhabitants, diameter, network opr 1, footprint building, PVC_medium			Total length, pollution, inhabitants, diameter, network opr 1, footprint building, PVC_medium, Percentage trees		
Lambda	**0.1**	**0.25**	**0.5**	**0.1**	**0.25**	**0.5**	**0.1**	**0.25**	**0.5**	**0.1**	**0.25**	**0.5**
train avg	0.56	0.52	0.49	0.56	0.52	0.49	0.57	0.53	0.50	0.58	0.54	0.51
test min	-0.07	-0.11	-0.16	-0.07	-0.11	-0.16	-0.31	-0.32	-0.36	-0.31	-0.33	-0.37
test max	0.62	0.60	0.60	0.63	0.60	0.60	0.62	0.61	0.60	0.57	0.57	0.56
Test AVG	**0.36**	**0.33**	**0.29**	**0.36**	**0.33**	**0.29**	**0.28**	**0.25**	**0.22**	**0.25**	**0.22**	**0.19**

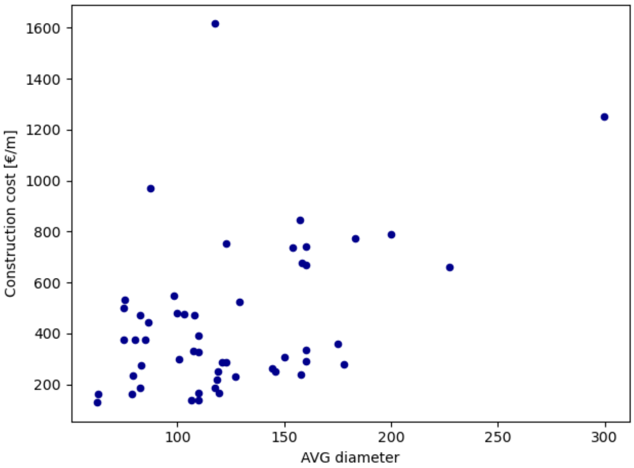


Figure C.1: Scatterplot average diameter gas projects

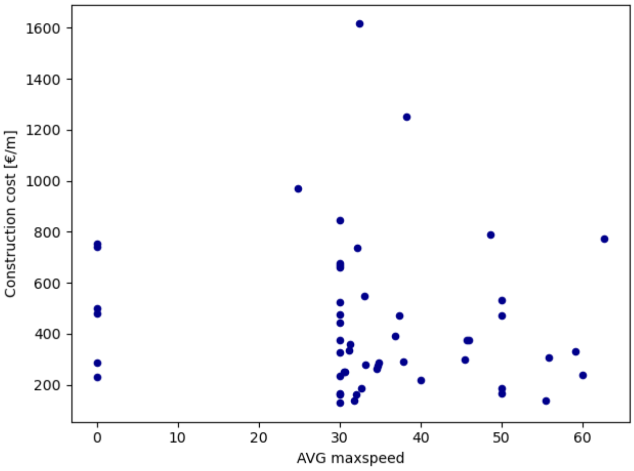


Figure C.2: Scatterplot average max speed gas projects

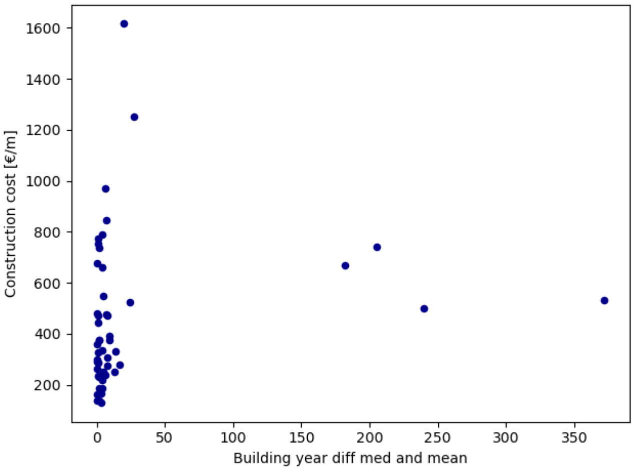


Figure C.3: Scatterplot buidling year differance mediaan and mean gas projects

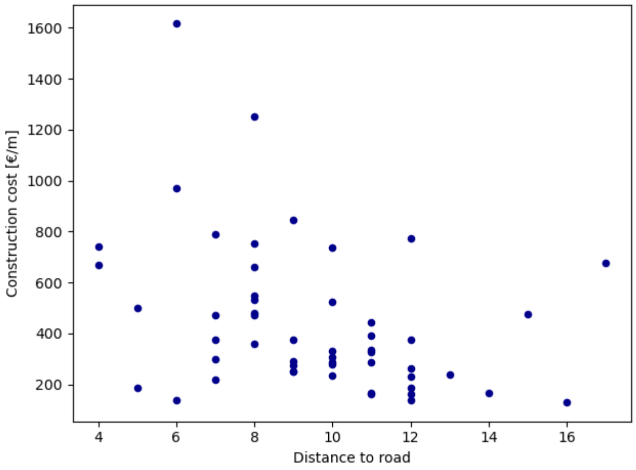


Figure C.4: Scatterplot average distance to road gas projects

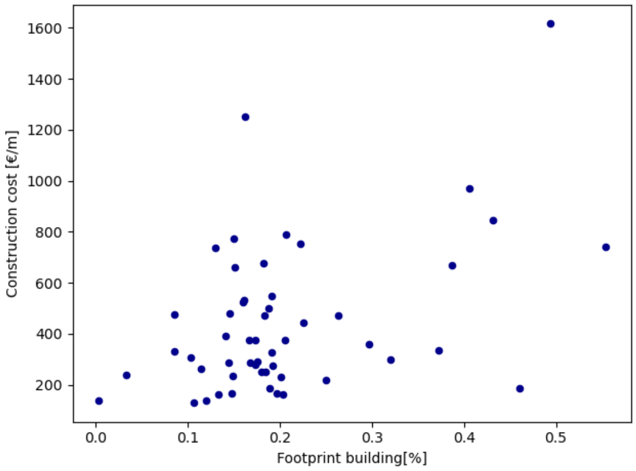


Figure C.5: Scatterplot footprint road gas projects

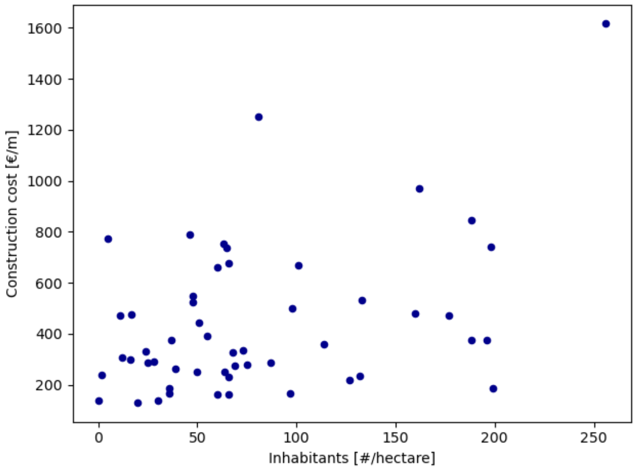


Figure C.6: Scatterplot inhabitants gas projects

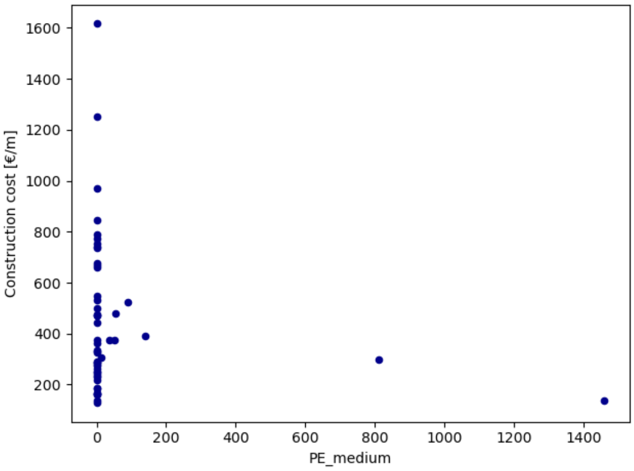


Figure C.7: Scatterplot PE medium gas projects

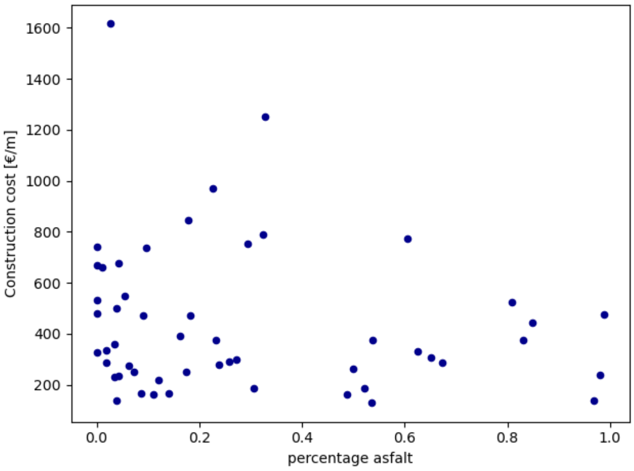


Figure C.8: Scatterplot percentage asphalt gas projects

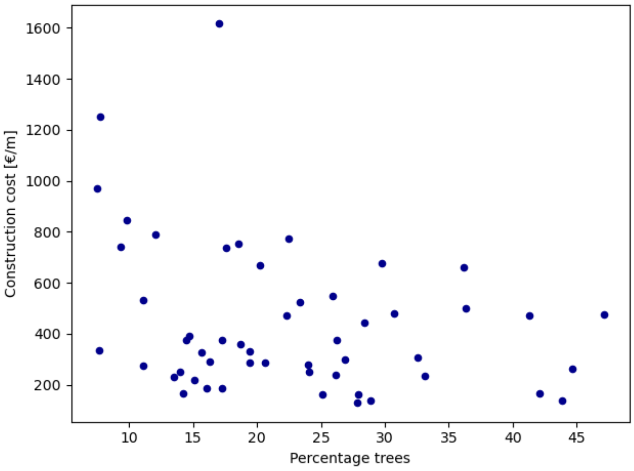


Figure C.9: Scatterplot percentage trees gas projects

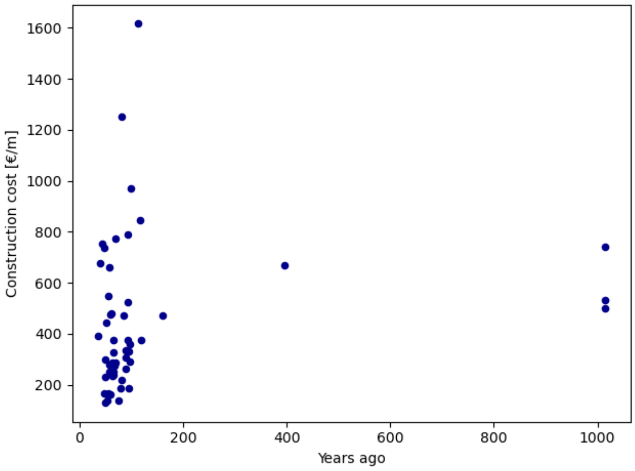


Figure C.10: Scatterplot years ago gas projects

D

Extra results Water

Table D.1: Model scores water model, trying different lambda's for L2/L1 regularization

Model settings: n_kfold=6, Regularization=True, reg_L1_wt = 0.5, Bagging=False

Inputs model:	Diameter, footprint building, Water bedrijf 1, road overig, Pollution category, Water bedrijf 2, Dewatering			Diameter, footprint building, Water bedrijf 1, road overig, Pollution category, Water bedrijf 2, Dewatering, Planned time			Diameter, Total length, Footprint building, Road overig, Dewatering, Water bedrijf 1, Pollutioin category, Water bedrijf 2			Diameter, footprint building, Water bedrijf 1, Road overig, Pollution category, Water bedrijf 2		
Lambda	**0.1**	**0.25**	**0.5**	**0.1**	**0.25**	**0.5**	**0.1**	**0.25**	**0.5**	**0.1**	**0.25**	**0.5**
train avg	0.60	0.57	0.53	0.60	0.56	0.53	0.61	0.58	0.55	0.58	0.55	0.51
test min	-0.19	-0.02	-0.11	-0.38	-0.42	-0.50	-0.41	-0.26	-0.24	-0.15	-0.02	-0.11
test max	0.49	0.52	0.54	0.49	0.51	0.52	0.50	0.53	0.56	0.47	0.50	0.52
Test AVG	**0.21**	**0.22**	**0.20**	**0.15**	**0.12**	**0.10**	**0.22**	**0.23**	**0.20**	**0.22**	**0.22**	**0.19**

Table D.2: Model scores water model, trying different lambda's for L1 regularization

Model settings: n_kfold=6, Regularization=True, reg_L1_wt = 0, Bagging=False

Inputs model:	Diameter, footprint building, Water bedrijf 1, road overig, Pollution category, Water bedrijf 2, Dewatering			Diameter, footprint building, Water bedrijf 1, road overig, Pollution category, Water bedrijf 2, Dewatering, Planned time			Diameter, Total length, Footprint building, Road overig, Dewatering, Water bedrijf 1, Pollutioin category, Water bedrijf 2			Diameter, footprint building, Water bedrijf 1, Road overig, Pollution category, Water bedrijf 2		
Lambda	**1**	**3**	**6**	**1**	**3**	**6**	**1**	**3**	**6**	**1**	**3**	**6**
train avg	0.61	0.59	0.56	0.64	0.59	0.55	0.60	0.59	0.55	0.58	0.56	0.54
test min	-0.46	-0.34	-0.34	-0.71	-0.42	-0.46	-0.57	-0.49	-0.53	-0.41	-0.35	-0.35
test max	0.46	0.47	0.48	0.48	0.46	0.49	0.45	0.46	0.46	0.42	0.44	0.46
Test AVG	**0.10**	**0.09**	**0.08**	**0.11**	**0.08**	**0.01**	**0.06**	**0.07**	**0.03**	**0.08**	**0.10**	**0.05**

Table D.3: Model scores water model, trying different number of models bagging

Model settings: n_kfold=6, Regularization=False, Bagging=True

Inputs model:	Diameter, footprint building, Water bedrijf 1, road overig, Pollution category, Water bedrijf 2, Dewatering			Diameter, footprint building, Water bedrijf 1, road overig, Pollution category, Water bedrijf 2, Dewatering, Planned time			Diameter, Total length, Footprint building, Road overig, Dewatering, Water bedrijf 1, Pollutioin category, Water bedrijf 2			Diameter, footprint building, Water bedrijf 1, Road overig, Pollution category, Water bedrijf 2		
n_models	**8**	**20**	**40**	**8**	**20**	**40**	**8**	**20**	**40**	**8**	**20**	**40**
train avg	0.65	0.65	0.65	0.66	0.66	0.66	0.66	0.66	0.66	0.65	0.65	0.65
test min	-0.93	-0.88	-0.90	-0.67	-0.57	-0.60	-1.15	-1.11	-1.11	-1.07	-0.96	-0.94
test max	0.51	0.49	0.52	0.55	0.53	0.55	0.52	0.51	0.53	0.50	0.48	0.51
Test AVG	**0.09**	**0.11**	**0.16**	**0.11**	**0.11**	**0.18**	**0.06**	**0.08**	**0.14**	**0.05**	**0.09**	**0.15**

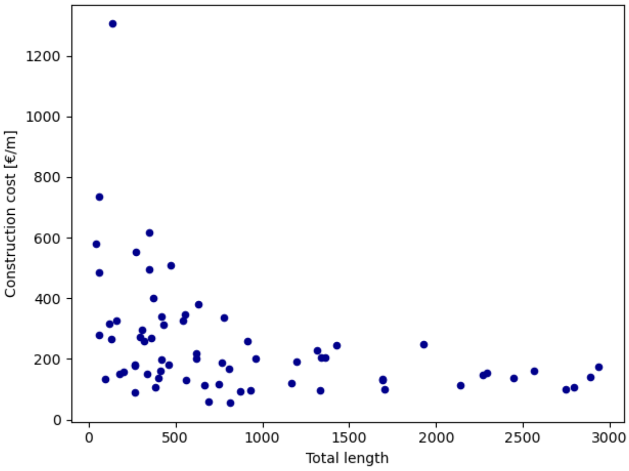


Figure D.1: Scatterplot total length water projects

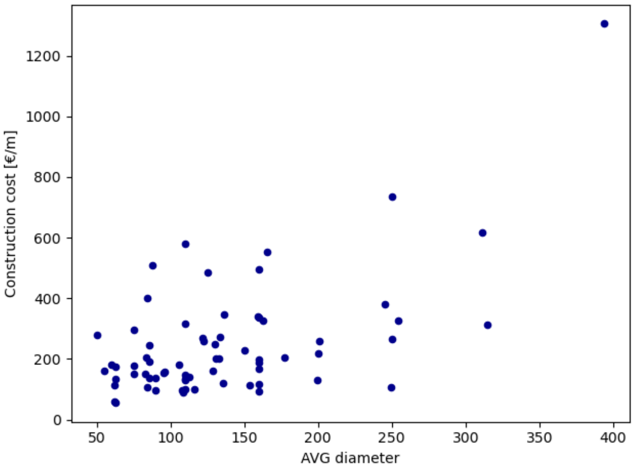


Figure D.2: Scatterplot average diameter water projects

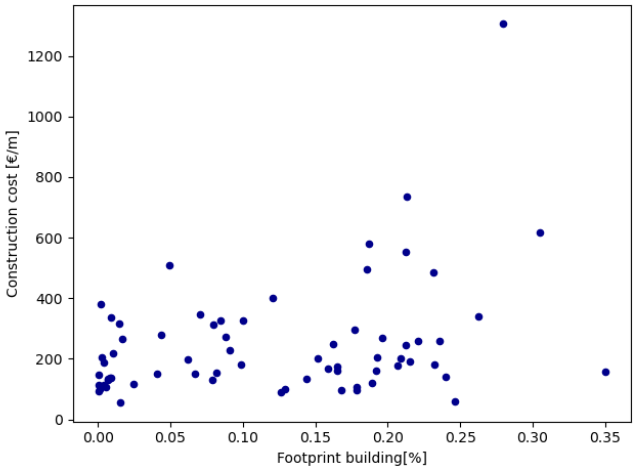


Figure D.3: Scatterplot footprint building water projects

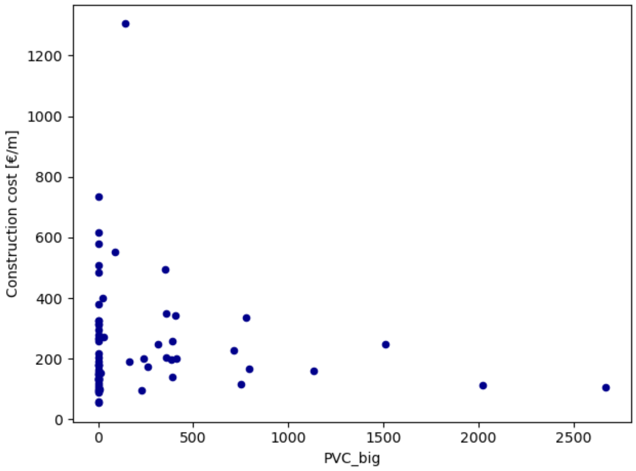


Figure D.4: Scatterplot PVC big water projects

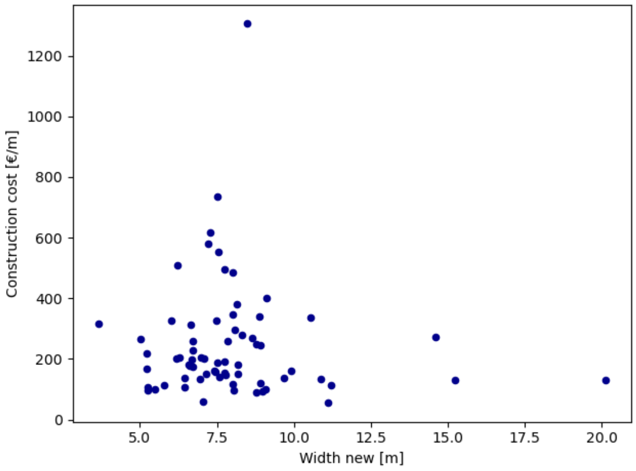


Figure D.5: Scatterplot road width water projects