



ASPARI

UNIVERSITY OF TWENTE

Bachelor Thesis Civil Engineering

**DEVELOPING A DATA
INFRASTRUCTURE
FOR AUTOMATING THE ASPHALT
COOLING PROCESS ANALYSIS**

PREPARED BY:

Grace Mora Ordoñez

SUPERVISED BY:

Dr. Seirgei Miller

Ir. Qinshuo Shen

Ir. Jasper Keizer

AUGUST 07, 2023

Bachelor Thesis in Civil Engineering

Author: Grace Anabel Mora Ordoñez
Student number: s2210029
Date: 07-08-2023
Version: Final Report

Student information

University of Twente
Grace Anabel Mora Ordoñez
g.a.moraordonez@student.utwente.nl

External supervisor

Adv. O&A Wegen
Ir. Jasper Keizer
jkeizer@kws.nl

Internal supervisor

University of Twente
dr. Seirgei.R. Miller
s.r.miller@utwente.nl

Daily internal supervisor

University of Twente
Ir. Qinshuo Shen
q.shen@utwente.nl

Preface

This report marks the culmination of my four-year study journey at the University of Twente to obtain my Bachelor of Science degree in Civil Engineering. As you will see, this is the final product of my bachelor thesis work, conducted from May to July in collaboration with ASPARi. The project is titled “Developing a Data Infrastructure for Automating the asphalt cooling process analysis.” The primary goal was to provide a data infrastructure for the PQi framework that automates the collection, storage, integration, processing, and analysis related to the asphalt cooling process.

I would like to thank and express my appreciation and gratitude to my supervisors. Dr. Sergei Miller for giving me the opportunity and encouraging me to work on this assignment, which has expanded my knowledge in different fields. To Ir. Quinshuo Shen for his constant guidance, support, patience, and positive attitude throughout the process. To Ir. Jasper Keizer from KWS, for accepting the invitation to join and supervise this thesis and all his time expended.

Apart from academics, I would like to also to thank my family and friends. Especially to my unconditional supporters during this entire journey my sister Cyntia Mora, my father Mesias Mora, and my mother, Carmen Ordoñez. Also, to my friends Liran, Snighda, and Coen for their advice and backing. Finally, I want to express my gratitude to the Ecuadorian government and the Secretaría de Educación Superior, Ciencia, Tecnología e Innovación - SENESCYT- as a sponsoring entity of my study program.

Grace Mora

July 2023

Summary

The construction process of asphalt pavement is divided into four phases: production, transportation, paving, and compaction. All these phases play a role in determining the quality and durability of the asphalt. However, the most crucial is the last phase -compaction-. Here the asphalt is compacted until a pre-determined density is reached. To ensure optimal compaction, operating within a specific temperature range known as the compaction window is essential. Therefore, it is important to acknowledge that the cooling behavior of asphalt during compaction will be influenced by several internal and/or external factors such as ambient temperature, wind speed, mix temperature at delivery, the temperature of subsurface, speed of pavers, roller capacity, type of mixture, and the temperature variation of the mix.

Since the compaction phase is often reliant on experience-based decision-making and craftsmanship of on-site operators, a certain level of variability and uncertainty is introduced, which may affect the quality of asphalt. In this context, ASPARi developed the Process Quality Improvement (PQi) framework that aims to provide better guidance for construction operations to transit from the standard support systems. The focus of this research is the cooling curve station of the PQi framework which has some limitations to be addressed.

Currently, the management of cooling curve data lacks adequate infrastructure, leading to ad hoc practices and fragmented processes. The proposed automated data infrastructure aims to overcome these limitations and provide a more organized and streamlined approach.

Therefore, this project aims to enhance the quality of data management in the PQi framework. This was done by developing and providing ASPARi with an automated data infrastructure to analyse the asphalt cooling behavior and its curve. It includes strategies for data collection, such as identifying the parameters and information that must be collected before and during construction, determining suitable and standard equipment, and implementing appropriate distribution and labeling of sensors. Also, a new database to store the collected information.

Moreover, a data preparation pipeline was designed using Extract, Transform, and Load (ETL) tools and various data cleansing methods to ensure the information's quality. Finally, a conceptual model was created. It estimates the asphalt cooling curve using the information from the database and polynomial regression.

The final product was subjected to expert opinion for validation and verification. This allowed the assessment of the developed data architecture and contributed valuable insights and recommendations for future research. The overall acceptance of methods, theories, completeness, and strategies was good.

Nederlandse Samenvatting

Het bouwproces van asfaltverharding is verdeeld in vier fasen: productie, transport, leggen en verdichting. Al deze fasen spelen een rol bij het bepalen van de kwaliteit en duurzaamheid van het asfalt. Echter, de meest cruciale fase is de laatste fase - de verdichting. Hier wordt het asfalt samengeperst tot een vooraf bepaalde dichtheid is bereikt. Om optimale verdichting te waarborgen, is het essentieel om binnen een specifiek temperatuurbereik te werken, bekend als het verdichtingsvenster. Daarom is het belangrijk om te erkennen dat het koelgedrag van het asfalt tijdens de verdichting wordt beïnvloed door verschillende interne en/of externe factoren, zoals omgevingstemperatuur, windsnelheid, temperatuur van het mengsel bij levering, temperatuur van de ondergrond, snelheid van de machines, capaciteit van de walsen, type mengsel en temperatuurvariatie van het mengsel.

Aangezien de verdichtingsfase vaak afhankelijk is van besluitvorming op basis van ervaring en vakmanschap van de operators ter plaatse, wordt er een zeker niveau van variabiliteit en onzekerheid geïntroduceerd, wat van invloed kan zijn op de kwaliteit van het asfalt. In deze context heeft ASPARi het Process Quality Improvement (PQi) framework ontwikkeld, dat tot doel heeft betere begeleiding te bieden voor bouwactiviteiten om over te stappen van de standaard ondersteuningssystemen. De focus van dit onderzoek ligt op het cooling curve station van het PQi-framework, dat enkele beperkingen heeft die moeten worden aangepakt.

Momenteel ontbreekt het aan een adequate infrastructuur voor het beheer van cooling curve data, wat leidt tot ad hoc praktijken en gefragmenteerde processen. De voorgestelde geautomatiseerde data-infrastructuur heeft tot doel deze beperkingen te overwinnen en een meer georganiseerde en gestroomlijnde aanpak te bieden.

Daarom heeft dit project tot doel de kwaliteit van het databeheer in het PQi-framework te verbeteren. Dit werd gedaan door het ontwikkelen en leveren van een geautomatiseerde data-infrastructuur aan ASPARi om het koelgedrag van het asfalt en de curve te analyseren. Het omvat strategieën voor gegevensverzameling, zoals het identificeren van de parameters en informatie die vóór en tijdens de bouw moeten worden verzameld, het bepalen van geschikte en standaardapparatuur, en het implementeren van geschikte verdeling en labeling van sensoren. Ook omvat het een nieuwe database om de verzamelde informatie op te slaan.

Bovendien werd een gegevensvoorbereidingspijplijn ontworpen met behulp van Extract, Transform, and Load (ETL) tools en verschillende methoden voor gegevensreiniging om de kwaliteit van de informatie te waarborgen. Ten slotte werd een conceptueel model gemaakt. Het schat de asfaltkoelcurve met behulp van de informatie uit de database en polynomiale regressie.

Het uiteindelijke product werd onderworpen aan expertbeoordeling voor validatie en verificatie. Dit maakte de beoordeling van de ontwikkelde data-architectuur mogelijk en leverde waardevolle inzichten en aanbevelingen op voor toekomstig onderzoek. De algehele acceptatie van methoden, theorieën, volledigheid en strategieën was goed.

Contents

Preface	iii
Summary	iv
Nederlandse Samenvatting	v
Table of Figures	vii
Table of tables	vii
1. Introduction	1
1.1. Problem statement.....	2
1.2. Research objectives and questions.....	3
1.3. Scope of the research	4
2. Methodology.....	5
2.2. Problem investigation.....	6
2.3. Treatment design	7
2.4. Treatment validation	9
3. Problem investigation	10
3.1. Asphalt cooling curve.....	10
3.2. PQi framework.....	11
3.3. Data management	13
3.4. Stakeholder Analysis.....	15
4. Treatment design.....	17
4.1. System functionalities.....	17
4.2. Data Collection	18
4.3. Data preparation pipeline.....	23
4.4. Data Analysis Mechanism	28
5. Treatment validation.....	29
5.1. Case study.....	29
5.2. Expert Opinion.....	31
6. Discussion	32
7. Conclusions and future work	35
7.1 Future work	36
8. References	37
Annex	40
A. ETL framework and process	40
B. Extra features of the system	41
C. Python libraries and scrips description	43

D. PQi relational database	45
E. RMSE and R2	45
F. Expert Opinion Session.....	46
G. Case Study	48

Table of Figures

Figure 1. Pavement node setup (Makarov et al., 2021).	2
Figure 2. Cooling of the asphalt mixture and compaction window (Miller et al., 2019)	3
Figure 3. Design Cycle Methodology. Adapted from Wieringa, 2014	5
Figure 4. Design process and its internal tasks.....	6
Figure 5. Representation of the literature review process (Quinshuo, 2023).....	7
Figure 6. Function tree overview.....	8
Figure 7.PQi-framework overview (Makarov et al., 2021)	12
Figure 8. Data management process.....	13
Figure 9. The general framework for ETL processes	14
Figure 10. Function tree overview.....	17
Figure 11. Set up of the asphalt node [adapted from (Makarov et al., 2021)].....	23
Figure 12. Snowflake scheme of the relational database.....	24
Figure 13. IR camera temperatures plotted including the regression and std dev	26
Figure 14. Identifying and correcting outliers	26
Figure 15. Missing data points	27
Figure 16. Asphalt cooling curve- dummy data.....	28
Figure 17. Raw IR camera data.....	29
Figure 18. After transformation IR camera data	29
Figure 19. After transformation surface temperature	30
Figure 20. Asphalt colling curve MP 1	30
Figure 21. Pedigree matrix results.....	32

Table of tables

Table 1. Factors influencing the cooling curve.....	11
Table 2. Overview of the stakeholders and their needs	16
Table 3. Information to be collected.	19
Table 4. Equipment used for data collection.....	19
Table 5. RMSE values of the regressions	28
Table 6. Pedigree scores.....	47

1. Introduction

To ensure the quality of asphalt, it is essential to pay careful attention to the construction process since this will determine the functional and mechanical properties of the material (Bijleveld, 2015). According to Bijleveld (2015), compaction is the last phase of the construction process, and during this phase, the asphalt must be correctly compacted to achieve a predetermined density. Moreover, Makarov et al. (2021) mention the influence of the design, execution, logistics, and environmental factors during this phase. Therefore, since any operational interruption or variability during this phase can significantly impact the quality and service life of the final product, this is considered a key phase that has a significant influence on the strength and durability of the asphalt pavement.

An essential factor during the asphalt compaction phase is the temperature of the asphalt mixture at which compaction is performed (Juma, 2020; van Deer, 1999). On the one hand, if compaction is done at a too high temperature, the binder is too fluid and causes the material to be displaced or shoved off due to low resistance against deformation. On the other hand, low temperatures increase the mix resistance resulting in poor lubrication, little compaction, and thus, low density (Juma, 2020; van Deer, 1999). Therefore, an optimal compaction time frame called the compaction window is specified during the mixture design (Makarov et al., 2021). This window is represented by two points (see Figure 20), the first is the highest temperature and time when the compaction should begin, and the second is the lowest temperature and time that points out that the passes of rollers are not optimal anymore (Ayuquina, 2022). The intended design properties will be accomplished if compaction is carried out during this frame (Bijleveld, 2015). Besides, there are also several parameters to consider for the compaction of asphalt, such as ambient temperature, wind speed, mix temperature at delivery, the temperature of subsurface, speed of pavers, roller capacity, type of mixture, and the temperature variation of the mix (Makarov et al., 2021; Arbeider, 2017).

However, due to the high reliance on intuitive and experience-based decision-making and craftsmanship, the construction process and associated operations are carried out by expert operators and managers with knowledge and experience. For instance, the roller operator must ensure optimal compaction effort is made during the compaction window. Thus, it is argued that the quality of the compaction phase is highly dependent on these people's expertise which causes variability and uncertainty (Makarov et al., 2021). To address this issue, the Process Quality improvement (PQi) framework has been developed to transit from a support to a guidance system that provides prescriptions and suggests actions/strategies to consider in the construction phase (Makarov et al., 2021). More specifically, this framework collects, stores, and analyses the data related to temperature and compaction in real-time and transforms it into actual strategies that operators can use to ensure the pavement's quality.

Given the importance of adequate compaction, it is therefore essential to ensure that the PQi framework and the associate Operator Support System (OSS) can provide correct guidance

and feedback regarding the cooling process of the asphalt pavement using the available equipment, to correctly indicate the current temperature of the asphalt to improve the decision-making concerning whether the asphalt is too hot, too cold, or adequate to be compacted. To perform this, the asphalt node in the PQi framework functions as the pivotal module to provide real-time measurement of the surface and core temperature of the asphalt layer and stream the data to the processor to generate the cooling curve of the asphalt pavement, thus providing a profile of asphalt cooling behavior.

This paper presents a new purpose for an automated and more organized data infrastructure to be used in the cooling curve station of the PQi framework. The focus is to facilitate the estimation and analysis of the asphalt cooling curve. Therefore, it is organized into five chapters. The first chapter details the problem to be solved and the objectives and questions related to the project. The second chapter presents all the relevant background information about the asphalt cooling curve and the PQi framework. Next, the third chapter shows the methodology followed to achieve the objectives. The fourth chapter is about the results and discussion of the model and the experts' opinions. The last chapter presents the conclusion and recommendations for future research.

1.1. Problem statement

The PQi framework is structured into three steps: data collection, data processing, and data analysis. The data is collected and transmitted using an IoT (Internet of Things) concept. Using sensor networking, all the data is collected at the asphalt, paver, and roller nodes. For this project, the focus is on the cooling curve module of the framework. Therefore, the first node -asphalt- is crucial because it measures, collects, and transmits the core temperature at different depths of the asphalt profile. As seen in Figure 1, this node measures the temperature using a set of thermocouples and a thermologger to collect and transmit the information (Makarov et al., 2021).

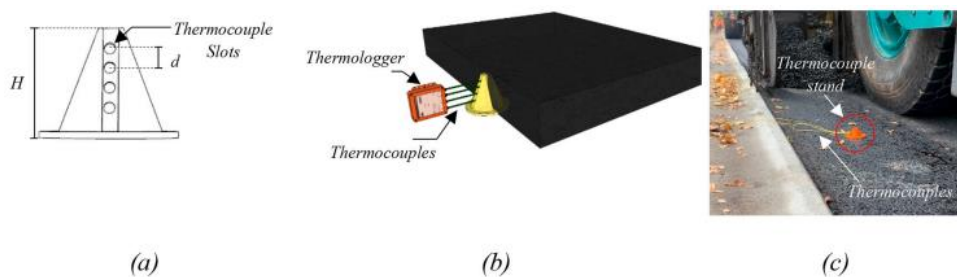


Figure 1. Pavement node setup (Makarov et al., 2021).

The processing node then uses this data to generate the cooling curve, as seen in Figure 2. This cooling curve offers a basis for determining the compaction window or amount of time left to complete compaction operations on-site using $\Delta t = t_1 - t_0$. This window is based on a predetermined temperature [°C] range for compaction specific to the asphalt mixture [T0, T1] (Makarov et al., 2021). Therefore, the results will help to ensure the quality of asphalt since the operators can estimate the time that compaction must start, and the time left

before the temperature of asphalt is too low for compaction. As aforementioned, correct compaction determines the quality and characteristics of the asphalt. Thus, it is essential to have clear and accurate results from the cooling curve.

However, this framework has some limitations that need to be addressed. The main issue is that the asphalt node lacks effective data infrastructure. Currently, data management is treated in an ad hoc manner. The entire process must be repeated in each new project, from setting strategies to collecting data, dealing with missing information, and filling in coefficients. This means there is no clear plan or consistent strategy for the acquisition, integration, storage, processing, and analysis of the asphalt cooling data. Moreover, there is fragmentation between the different stages of the process. For instance, for some information not correctly transmitted by the nodes, the operators must include themselves, such as missing temperature readings and correct magnitudes. Also, the processing and analysis of the cooling curve still rely on handwritten documents, which pushes the operators to include this information by hand in the program. This fragmentation consumes much more time and hinders the further archiving, analytical, and modelling processes. Therefore, incorporating a proper data infrastructure can advance the regression of the cooling curve by using techniques such as machine learning.

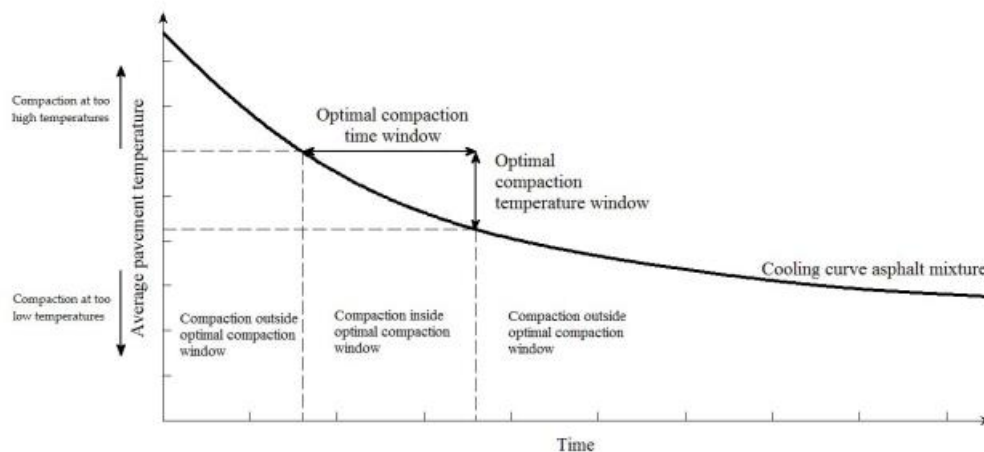


Figure 2. Cooling of the asphalt mixture and compaction window (Miller et al., 2019)

1.2. Research objectives and questions

A new data infrastructure needs to be designed to have a more effective framework. This infrastructure should automate the data collection, transmission, and analysis of the data. Hence, it is necessary to transition from a case-by-case basis to a more encompassing strategy that organizes the information depending on its purpose while keeping quality and accuracy. Furthermore, the new infrastructure should provide more analytical and relevant outcomes for post-assessments and enable efficient and effective data sharing among organizations. Therefore, the objective of the research is formulated as follows:

To develop a data infrastructure to automate the collection, storage, integration, processing, and analysis of the asphalt cooling process and data so that it can facilitate the cooling curve prediction.

To achieve this objective, this main question was proposed:

- How can the information acquired by the PQi framework be managed to obtain more effective and efficient data processing, analysis, and visualization of the cooling behaviour of asphalt?

This main question will be answered by the following sub-questions:

- What data are needed to comprehensively demonstrate the cooling behavior of the asphalt?
- How to efficiently and effectively structure the data to automate the prediction of the asphalt cooling curve?
- How to collect corresponding data strategically on the construction site to cope with the project specification?
- How to automate the processes of data collection to data storage and the cooling curve data processing and analysis?
- How can the data be used by the ASPARi companies to improve the compaction strategies?

1.3. Scope of the research

This project aims to improve the data management of the PQi framework by developing a data infrastructure that automatizes all core processes to analyse and improve the understanding of the cooling behaviour of asphalt. It is important to mention that the framework works with three modules named cooling curve, temperature contour plot, and compaction contour plot. However, the cooling curve module will be the only focus for this project due to the time and complexity constraints.

For the current project, developing a data infrastructure implies providing a scheme or structured system that supports data management throughout its lifecycle. The data infrastructure will include several tools, techniques, and processes that reinforce the collection, preparation, analysis, and use of data for the estimation of the asphalt cooling curve. The purpose of the estimation is to support the post-construction process and analysis of the pavement and to cover the paved and compacted areas where no direct temperature cooling data has been collected to enable the application of OSS (Operator system strategies).

Additionally, the information that will be considered relevant data for the data infrastructure and the study will be of two types. First, general information refers to the fundamental details about the project, including location, client, contractors, mixture details, layers details, and equipment. Second, parameters that can be measured or collected at the asphalt node.

These are the temperatures of the core and surface layers, density progression, roller passes, ambient temperature, wind speed, humidity, pressure, and type of compaction.

2. Methodology

To proceed with the design of the data infrastructure for the estimation of the cooling curve in the PQi framework, the method selected is the design cycle. This method divides the design process into four tasks named problem investigation, treatment design, treatment validation, and system implementation. It allows iterations over the investigation and design processes that are involved in the project (Wieringa, 2014).

The current project will be focused mainly on the first three stages of the design cycle. First, investigating and understanding the problem by reviewing relevant literature concerning the PQi framework and data management. Additionally, a stakeholder analysis is included to understand the social context. Second, the treatment design includes the design requirements based on the previous phase, also the data collection, preparation, and analysis mechanisms. Third, the design or system is validated using experts' opinion and a case study to assess the performance of the designed data infrastructure.

It is important to mention that the last task of the system design was excluded since it involves the implementation of the system into the real PQi framework. This is hardly possible due to the time constraints and the complexity of combining this design with the actual PQi framework. It would require more resources, time, and expertise to ensure optimal results. Despite these limitations, the potential scenarios and possible implications of the implementation of the data infrastructure are discussed in Chapter 6. Figure 3 shows the overview of the cycle and its specifications.

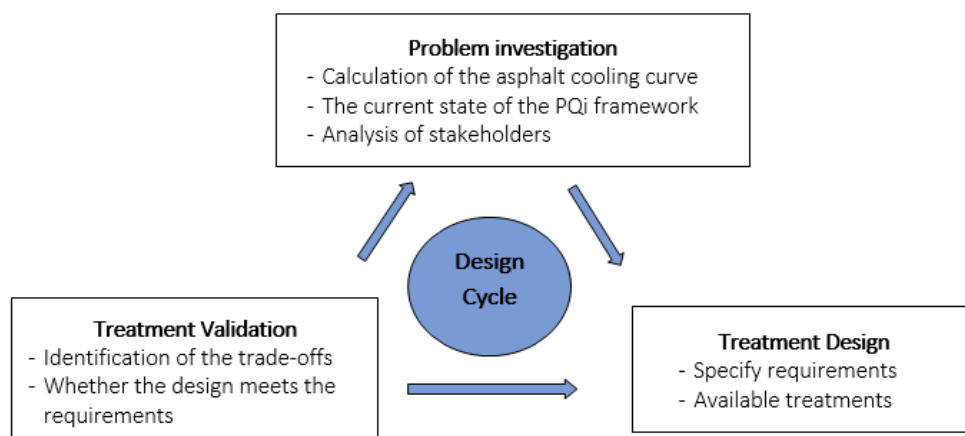


Figure 3. Design Cycle Methodology. Adapted from Wieringa, 2014

Using this cycle allowed the designer to establish the specific tasks for each phase of the design and track the desired outputs. Figure 4 shows the overview of the design process followed to develop the final product, and it reflects the relationship between different tasks. The following sections of this chapter include more details about each stage of the methodology of this project.

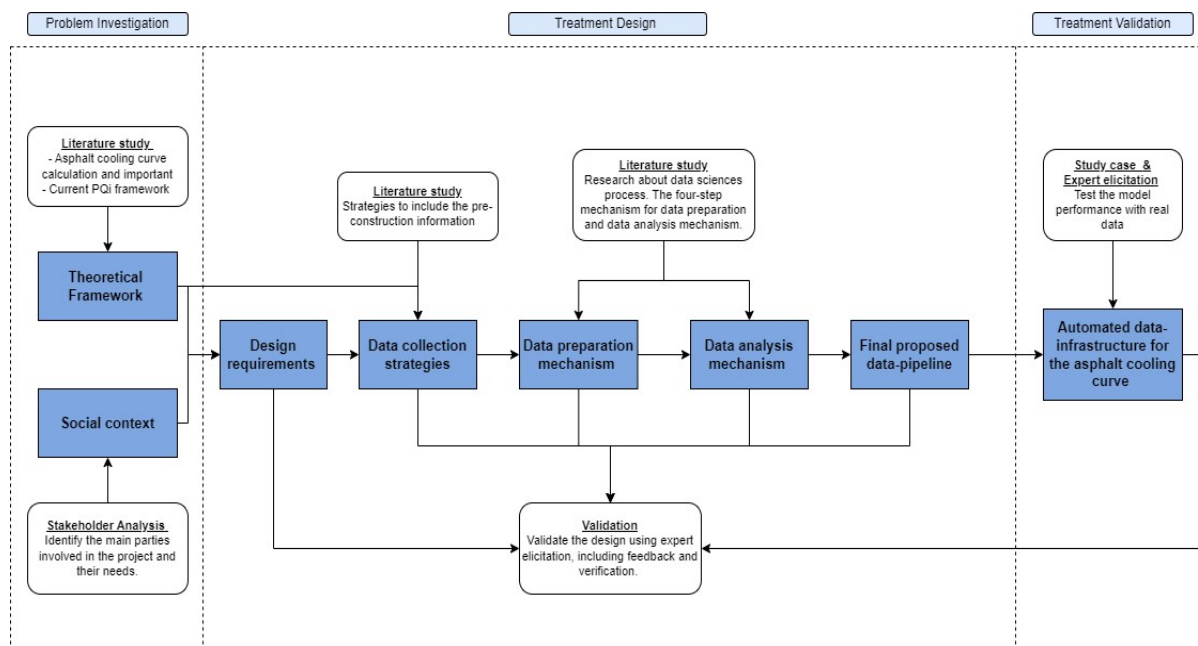


Figure 4. Design process and its internal tasks

2.2. Problem investigation

This is the first phase of the design process, and it includes two activities: the theoretical framework and the social context, both used to build up the design. The purpose of performing these activities is to explore, represent, and understand not only the problem itself but elaborate the theories that will be useful and support the system design phase.

2.2.1. The knowledge framework exploration

The theoretical framework provides the relevant knowledge for this design project through a critical examination of pertinent literature. The examination included key concepts and theories about the importance and the procedure used to calculate the cooling curve of asphalt, the current PQi framework, and data management procedures.

Therefore, the literature review was done based on the systematic process proposed by Onwuegbuzie and Frels (2016) seen in Figure 5. It consists of three phases named exploration, interpretation, and communication. For the first phase, it was necessary to establish the search engines and the keywords to be used for the literature exploration. The early group included Google Scholar, Lisa UT, and ResearcherGate, due to their vast and varied literature in the field. The keywords based on the topics of interest were “PQi framework,” “Asphalt cooling behaviour,” “Asphalt cooling curve,” “Data management,” “Data cleansing,” “Outliers identification,” “Missing data management,” etc. Based on this research, the information was filtered and stored to continue with the next phase – interpretation. Here, the information was analysed and summarised to proceed to the last part of structuring and organizing it coherently.

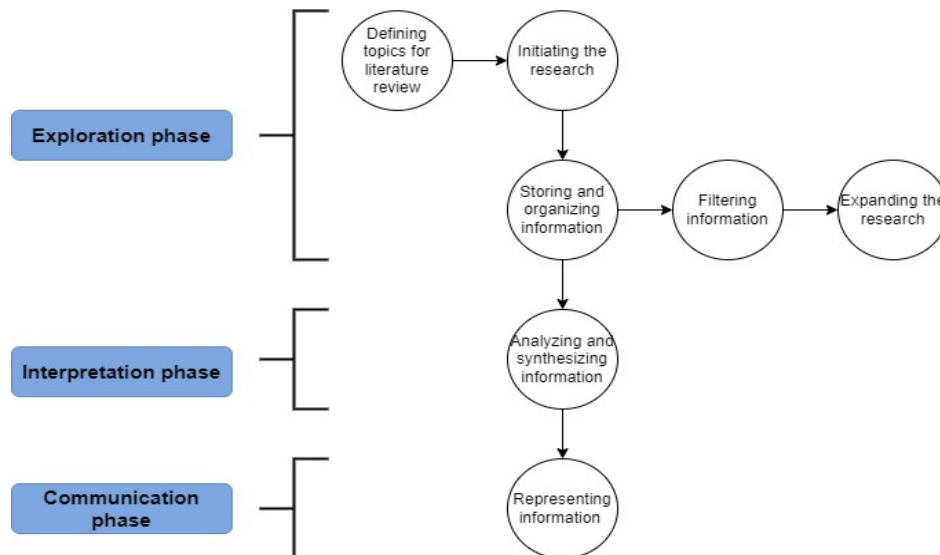


Figure 5. Representation of the literature review process (Shen, 2023)

2.2.2. The social context exploration

To explore the social context of the proposed system, a stakeholder analysis was performed. Examining the stakeholders in a project allows for revealing the actions, motives, relationships, agendas, interests, and resources of the participants (Coetzee et al., 2020). Therefore, this analysis started with the identification of all parties that hold a stake or share interest in the design. This was done during internal brainstorming sessions between the supervisor and the designer to understand for whom the design was made. Then, their needs and interest were discussed and determined. These are also the starting points for the system requirements.

2.3. Treatment design

This phase is focused on the actual development of the system by using the reviewed theory to establish first the data infrastructure requirements. Then, both explanations of the asphalt cooling curve calculation and the PQi framework were used to establish the relational database. Lastly, the data pipeline was proposed based on the data preparation and data analysis mechanisms which are further explained below.

2.3.1. Design functionalities

The first task in the treatment design phase was specifying the requirements that the data infrastructure should meet based on the desires and necessities of the stakeholders. Since the project consist of designing the data infrastructure for the cooling curve station, its requirements were specified in terms of the function requirements (FR). According to Kang (2010), FR is a function for which the designer defines the performance of the designed system. The hierarchy of the function requirements and the relationship among the functions of the system can be represented using a function tree. As seen in Figure 6, the left side of the tree represents what the system must achieve, and on the right, it shows how the designer wants to achieve the function requirement.

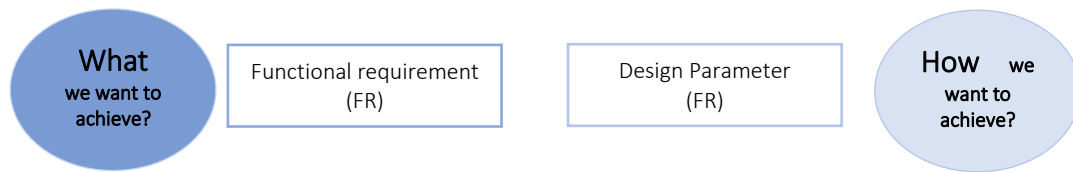


Figure 6. Function tree overview

2.3.2. Data collection strategies

The collection of data is an important stage since the quality and accuracy of the cooling curve and future analysis will highly depend on the quality of the acquired data. Thus, to have more efficient data collection, some strategies were proposed based on the objectives of the project. These included the classification of the information and/or parameters to acquire based on the source of the data, either the client or ASPARi. Also, the inventory of the ASPARi equipment that has been used for data collection contains the features and methods to ensure accurate measurements. Additional specifications are also covered, such as sensors placement along the area to be paved, time steps, and equidistance distribution of measurement points.

2.3.3. Data preparation pipeline

The preparation of the pipeline was done based on the system requirements and reviewed literature about data management. First, the database was created and organized based on the literature study and the multidimensional modelling method by Jansen et al. (2010) and Keulen & Ahmed (2021). This method consists of four steps, including the relational representation of the database using the snowflake scheme. The parameters were derived from the literature study on factors influencing the cooling curve of asphalt and the relational database employed by the PQi framework. Moreover, a PostgreSQL administration and development tool, pgAdmin 4, was the software employed in the creation of the database called "Asphalt cooling curve."

The second part was developing the pipeline for the data preparation. The ETL process (extract, transform, load) was used for this part. The sources of information were identified and loaded into a common data lake using Python. Subsequently, the transformation consisted in detecting and replacing outliers present in the dataset using polynomial regression. Additionally, the missing values present were resolved by using KNN imputation. This will be further explained in section 4.3. The final step of the pipeline is to upload the transformed data into the database using SQL language.

2.3.4. Data analysis mechanism

The data analysis mechanism is the method used to fulfil the model's objective, which is to estimate the asphalt cooling curve using the acquired data processed by the automated pipeline developed. This model is conceptual in nature and was developed using empirical data. Thus, a multivariable regression of degree two is used based on the collected and prepared data. The parameters considered were the core temperatures, ambient temperature, wind speed, pressure, and humidity. The final plot includes the asphalt cooling

curve, the *RMSE* and R^2 values that indicates the performance and the proportion of variance of the model.

2.4. Treatment validation

As part of the design cycle, the validation will be performed as the last step during each iteration. According to Wieringa (2014), the goal of this phase is to predict what would happen if the developed design were transferred to its problem context by simulating its implementation. Therefore, for each research output of the design cycle, a validation needs to be performed. The new inputs obtained from the validation phase will help to improve the design by identifying problems, finding the corresponding solutions, and ensuring the target requirements are achieved. To validate the developed data structure, a case study along with an expert session, during which experts will have the opportunity to share their valuable insights and opinions. These will allow experts to visualize how the model works and identify the problems and suggest improvements.

2.4.1. Case Study

A case study was carried out to test the performance of the developed data infrastructure. Therefore, real data that belongs to a project carried out on October 3rd, 2022, was provided for this matter. The objective of this study is to gather the documents, then perform the necessary preparation and transformations to get proper data for the estimation of the cooling curve.

2.4.2. Expert Opinion

By using this validation method, it is intended to submit the design to a panel of experts that understand or have knowledge about the subject (Weringa, 2014). This allows the researcher/designer to ask professionals for their views on the design's potential usefulness and usability (Weringa, 2013). Thus, experts visualize how the model works and identify the problems and suggest improvements. It is important to mention that even though positive opinions are valuable, these can be bias. Thus, negative or critical opinions are more useful since they can clear the researcher or designer's mind and provide more room for improvement (Weringa, 2014).

The objective of the expert opinion was to explore the final design using the expertise of the specialists to obtain their insights and feedback regarding the theoretical basis, methods, and performance of the designed data infrastructure. Thus, for this session, a PowerPoint presentation was prepared to explain to the experts the main information and concepts about the project and all the methods used for the data preparation, as well as the mechanism used to estimate the asphalt cooling curve. It also included a questionnaire and a pedigree matrix that can be found in Annex F.

The questionnaire sought to elicit insightful comments and inputs based on the experience of the experts and ensure coverage of the relevant aspects of the project, such as their overall impression, major gaps and areas of improvement, limitations, and weaknesses, usability of the system, possible challenges in the real world, and further improvements. Additionally, the

use of a pedigree matrix allowed the experts to qualitatively judge the data infrastructure based on the criteria on a numerical scale from 0 (weak) to 4 (strong) with specific descriptions of each level on the scale (van der Sluijs et al., 2004).

3. Problem investigation

This section will provide the background information to have a better understanding of the problem and the basic concepts and theories that will help to solve it. First, the process and parameters needed to calculate the asphalt cooling curve, as well as its importance, will be introduced. Then, the current PQi framework will be described. Subsequently, a description of how data needs to be managed is given. Finally, a stakeholder analysis is developed to identify the main parties involved in the project and their interests or needs.

3.1. Asphalt cooling curve

The asphalt cooling curve is a schematical representation of the cooling rate of the asphalt mixture (Bijleveld, 2007). The importance of this curve lies in the role that temperature plays during the compaction phase of the asphalt paving process. Both researchers and practitioners of the asphalt industry agree that the temperature of the asphalt mixture during compaction determines the quality of the road and its service life (Chadbourn et al., 1998; Arbeider, 2017; Youness, 2007). By carrying out proper compaction the desired asphalt density will be achieved. Some authors suggest a temperature range for compaction between 90 and 100 °C (Floss, 2001), while others suggest a maximum cut-off commonly at 130 °C (Commuri & Zaman, 2008) or minimum cut-offs at about 70 and 80 °C (van Dee, 1999), or 110 °C to obtain a good surface texture and acceptable relative compaction of the asphalt pavement mixture (Youness, 2007). Compaction done at different temperatures can affect positively or negatively the quality. If the temperature mixture is too high, the binder is too fluid that the rollers would displace the material rather than compact it and the material will be likely to crack. On the contrary, if the temperature is too low, the mixture is too viscous and difficult to work with. Also, the bitumen will not lubricate the mixture properly, which would result in an open surface with a tendency to raveling (Bijleveld, 2007). Furthermore, low temperatures also affect the resistance against deformation and the stiffness of the asphalt (Youness, 2007). Therefore, Timm et al. (2001) indicate that there is an optimal temperature range for compacting asphalt mixtures that will increase the likelihood that the appropriate mechanical properties will be accomplished. This implies that there is also an optimal time range for compacting, depending on the pace of cooling of the asphalt mixture (Bijleveld, 2007). Both ranges together as known as the optimal compaction window.

Figure 4 shows the cooling rate of an asphalt mixture as a function of time, and it also shows the optimal compaction window in which the asphalt mix must be compacted to ensure high quality. Two points are given, the first is the highest temperature and time when the compaction should begin, and the second is the lowest temperature and time that points out that the passes of rollers are not optimal anymore (Ayuquina, 2022). Moreover, it is important to mention that the ideal compaction window depends on the type of mixture and

the conditions under which the mixture has been compacted, and it shifts along the timescale.

Several factors influence the thermal behaviour of the asphalt layer during the construction stage, and it is crucial to consider all of them to properly estimate the time at which the rollers need to start compacting and the time left before the mixture is too cold for compacting (Makarov et al., 2021). Hence, Arbeider et al. (2017) categorize these facts influencing the cooling curve of the asphalt layer into three groups named mixture characteristics, weather conditions, and subsurface conditions. Table 1 displays all the factors that belong to each group, including some extra factors suggested by Vasenev et al. (2012).

Table 1. Factors influencing the cooling curve.

Mixture characteristics	Weather conditions	(Sub) Surface conditions
Type of mixture	Ambient temperature	Type of the existing surface
Binder properties	Wind speed	Material conditions
Delivery temperature	Cloud cover	Surface temperature
Layer thickness	Time of the day and year	Temperature of subsurface
Thermal conduction	Humidity	
Thermal transition coefficients	Pressure	

3.2. PQi framework

As aforementioned, to achieve high-quality roads, compaction must be completed within a specific temperature and time range, known as the optimal compaction window. If this is not done within that window, the mechanical properties of asphalt as well as the service life, are negatively affected. Nowadays, there are several areas of the construction industry, especially contractors of road infrastructure, that still rely noticeably on the on-site experience and tacit knowledge of operators and teams (Makarov et al., 2021). These people need to ensure compaction is done correctly. However, there is uncertainty in the results due to the high reliance on intuitive and experience-based decision-making and craftsmanship, which do not guarantee the high quality that contractors and clients look for.

To address this issue, the Process Quality improvement (PQi) framework has been developed to transit from a support to a guidance system that provides prescriptions and suggests actions/strategies to consider during the construction phase (Makarov et al., 2021). More specifically, this framework collects, stores, and analyses the data related to temperature and compaction in real-time and transforms it into actual strategies that operators can be used to ensure the quality of the pavement. The overview of this framework can be seen in Figure 7. Therefore, given the importance of effective compaction, it is essential to ensure that the PQi

framework and the associate Operator Support System (OSS) can provide correct guidance and feedback regarding the cooling process of the asphalt pavement to correctly indicate the current temperature of the asphalt to improve the decision-making concerning whether the asphalt is too hot, too cold, or adequate to be compacted. To perform this, the asphalt node in the PQi framework functions as the pivotal module to provide real-time measurement of the surface and core temperature of the asphalt layer and stream the data to the processor to generate the cooling curve of the asphalt pavement, thus providing a profile of asphalt cooling behavior.

However, as mentioned in section 1.1, the PQi-framework has some limitations that need to be addressed. The asphalt node's inadequate data infrastructure is the key problem. Currently, data management is handled arbitrarily, and each new project necessitates starting the procedure from scratch. This indicates that there is no defined strategy or plan in place for gathering, integrating, storing, processing, and analysing data on asphalt cooling behaviour. Additionally, the process is fragmented between its several phases. For example, the operators must include themselves approximated temperatures that the nodes are unable to convey correctly; in some cases, they also may remove certain readings or outliers that do not fit into the curve. Besides this, the operators are forced to manually enter this data into the program because some processes and analyses of the cooling curve still use handwritten papers, such as thermal coefficients. Due to this fragmentation, additional archival, analytical, and modelling operations take substantially longer and are hindered.

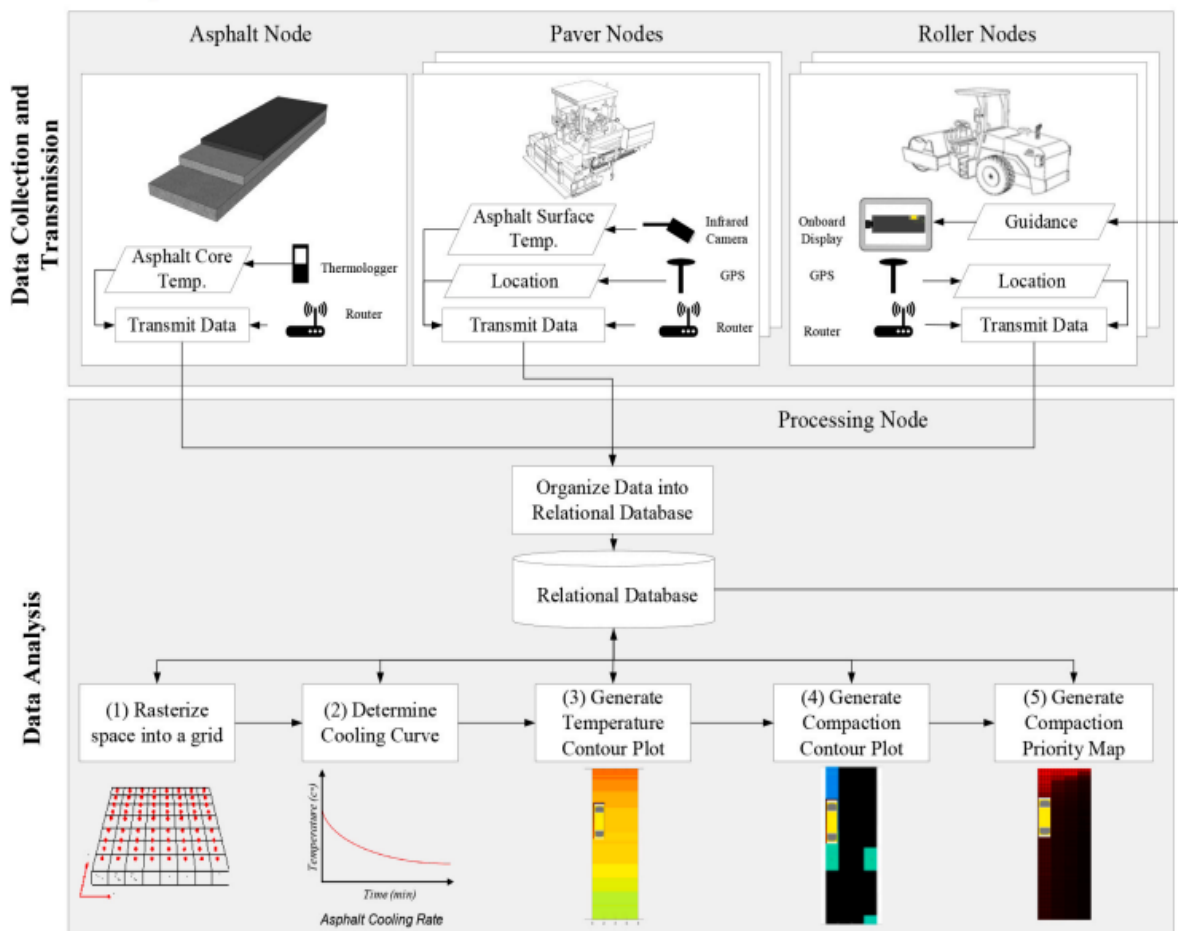


Figure 7. PQi-framework overview (Makarov et al., 2021)

3.3. Data management

To gain knowledge about the subject, it is important to investigate the concepts and mechanisms involved in multidimensional data management. The models developed using this approach are designed basically to support data analyses (Jansen et al., 2010). Figure 8 shows the process that needs to be followed to correctly work with the data and obtain the desired outputs. It consists of four steps named data collection, data preparation, data analysis, and use. Data scientists have reported spending 80% of their work time on data preparation, specifically cleaning (van Keulen & Ahmed, 2021).

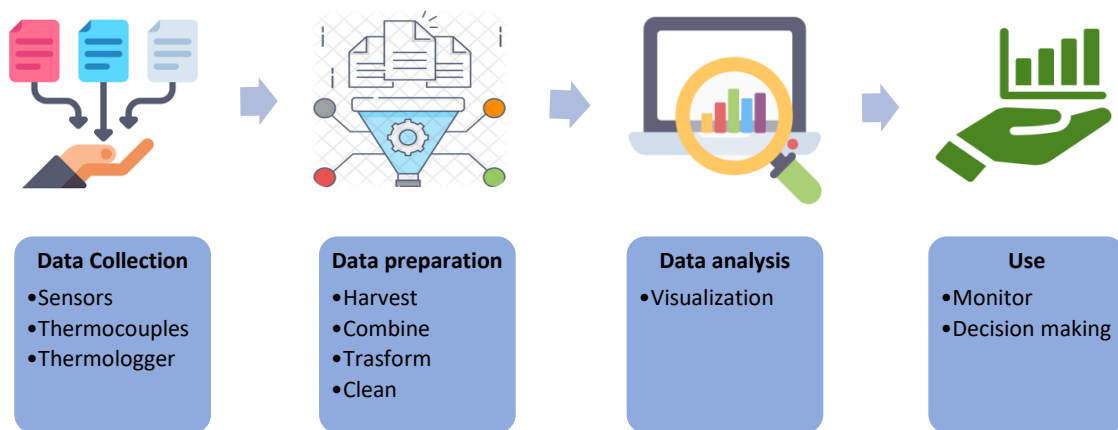


Figure 8. Data management process (van Keulen & Ahmed, 2021).

3.3.1. Multidimensional modelling data preparation

The data preparation will be organized based on the method for data preparation of the multidimensional modelling by Jansen et al. (2010) and Keulen & Ahmed (2021). This method consists of four steps, which will be listed and explained below.

1. Design a cube by using a snowflake scheme.
2. Design associated table structure.
3. Create (empty) tables in the database.
4. Prepare data and fill tables.

Before starting with the first step, it is crucial to understand the fundamental concepts of multidimensional databases: cubes, dimensions, facts, and measures. Thus, a cube is a term used to refer to multidimensional databases; it is able to capture and analyse data (Jansen et al., 2010). It contains dimensions, as much as the designer wants to provide all the necessary context for the final goal. Dimensions are meant to select and group the data at a certain level of detail (Jansen et al., 2010). Moreover, the cube also consists of unique cells that can be easily identified. When the cell is not empty, it is called a fact that represents the subject that is to be analysed. Each fact also can contain a numerical property or measure. The

snowflake scheme is the representation of the cube dimensions. It helps to introduce the hierarchy of the data to be used during the last stage of the data lifecycle.

Therefore, the first step for the data preparation consists of designing the relational representation of the data by using the snowflake schema. This will select and group the data into the different levels of detail that the designer considers necessary. Then, the table structure will be arranged to organize and save the corresponding data. The next step is to create empty tables in the database where all the information will be stored. The last step of this method is to prepare the data and fill the designed tables, but the preparation is the essential process to have representative and high-quality data to analyse. This can be done using ETL processing that facilitates the data flow and is explained below.

3.3.2. ETL

To construct proper databases, it is necessary to run the ETL tools (Shaker et al., 2011). The Extract-Transform-Load (ETL) system serves as the basis for databases (Kimball & Caserta, 2004). The ETL tools are software components responsible for three main tasks: a) the extraction of data from different sources, b) The customization and cleansing of that data, and last c) the Insertion or loading of the data into a data warehouse (Shaker et al., 2011; Vassiliadis et al., 2002). Using the ETL system is more than moving data from sources to a data warehouse, and it provides significant value to data by removing mistakes and correcting missing data, providing documented measures of confidence in data, capturing the data flow for safe storage, adjusting data several to be used together, structuring data to be usable by user tools (Kimball & Caserta, 2004).

ETL is a combination of process and technology that is considered complex since it consumes a significant percentage (80%) of the data warehouse development time (Shaker et al., 2011; Vassiliadis et al., 2002). The general framework for ETL processes can be seen in Figure 9. The bottom layer represents the data stores involved in the entire process. From the left to the right, there are the sources, the data staging area (DSA), and the data warehouse (DW). The top layer represents the three consecutive steps: extraction, transformation, and loading. More information can be found in Annex A.

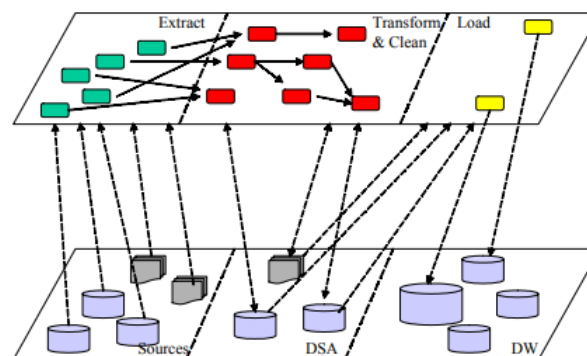


Figure 9. The general framework for ETL processes

3.4. Stakeholder Analysis

According to Freeman (1984), a stakeholder can be defined as any individual or group that is affected or affects the achievement of the project objective. Examining the stakeholders in a project allows for revealing the actions, motives, relationships, agendas, interests, and resources of the participants (Coetzee et al., 2020). Therefore, for the current design, it is important to identify all parties that hold a stake or share an interest in the design to understand for whom the design is made and what needs must be fulfilled in order to be successful and to create value for all actors involved.

Four stakeholders that influence and have power in the project were identified. These are described below, and Table 2 shows an overview of their needs.

- ASPARi: It is a knowledge network founded in 2007. It consists of researchers at the University of Twente, infrastructure contractors, and Rijkswaterstraat that work together to improve the asphalt-related construction processes in the Netherlands (ASPARi, n.d.). They have a high interest and influence in the project since the final design will be implemented into their PQi framework.
- Contractors: This group represents all the clients that will use the PQi framework at work and will benefit from the new automated data infrastructure during the on-site construction phase and using the available equipment. This group also includes the host company of the project - KWS - where the case study will be done. Thus, it has a high interest and medium influence.
- Supervisors: The project will be supervised by three persons who have vast knowledge in engineering and asphalt construction as well as data management. They will provide feedback and verified the design meets certain standards they consider necessary. Thus, high interest and high influence are considered for this group.
- Designer: This individual has a high interest and strong influence on the project since she is the researcher and developer of the design. All decisions should be made by her in consultation with the supervisors.

Table 2. Overview of the stakeholders and their needs

Stakeholders	Needs/requirements/role
ASPARI	<ul style="list-style-type: none"> • Improve the performance of the PQi framework by automating the internal process to give prescription guidance to the end users. • Transit from the ad-hoc to a more encompassing strategy for managing the acquired data during the asphalt pavement process. • Facilitated future inspections and analyses of the roads.
Contractors	<ul style="list-style-type: none"> • Prediction of the cooling curve use during the on-site asphalt construction • To save time during the use of the cooling curve prediction system and while interpreting the outputs. • To ensure better compaction guided by the cooling curve. • Appropriate performance of the framework to improve the compaction strategies during the construction phase of asphalt. • Enable collaboration between companies to facilitate information sharing and decision-making.
Supervisors	<ul style="list-style-type: none"> • Ensure the approach and path that the project takes are correct. • Establish the standards for the design and share feedback for possible improvements.
Designer	<ul style="list-style-type: none"> • Design an appropriate pipeline and data warehouse to automate the asphalt cooling analysis. • To obtain explicit knowledge related to data science and asphalt cooling behaviour. • Ensure the data infrastructure designed can efficiently and effectively store, retrieve, and process all the data to calculate the cooling curve.

4. Treatment design

This section presents all the details about the design of the new system. First, the functions of the system are stated. Second, all the strategies to improve the data collection are presented and described, which includes the exact information that needs to be collected, the equipment to be used and how to manage them, and the distribution of the sensors throughout the pavement area. Then, the data preparation is described, including the representation of the relational database and techniques to manage and cleanse the data. Finally, a description of how the actual cooling curve was calculated.

4.1. System functionalities

As mentioned in section (2.2.2), the requirements were specified based on the main functions the system must have to satisfy the stakeholders' needs. Figure 10 is the function tree, and it displays the hierarchy of these function requirements. This tree also represents the overall structure of the system.

The main function of the asphalt cooling curve structure is to provide effective and efficient data management by focusing on the three main aspects. First, strategies for the on-site data collection that integrates the pre-construction information. Second, an automatic data preparation pipeline that guarantees the quality of the data and reduces human intervention in the process. This will also reduce the inaccuracy during the analysis. Lastly, automatic data analysis generates the asphalt cooling curve.

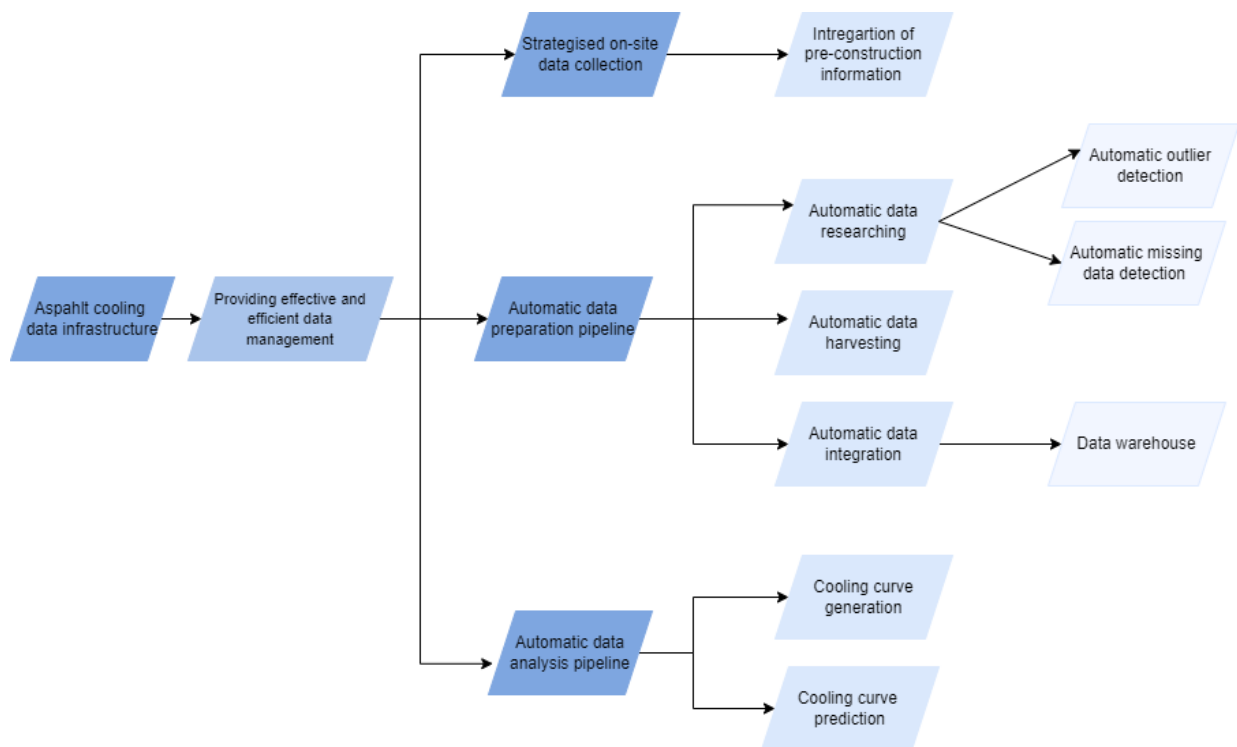


Figure 10. Function tree overview

4.2. Data Collection

The purpose of this section is to establish the strategies that will be used to perform better data collection that improves the performance of the cooling curve modulus of the PQi framework. The quality of the data acquired will determine the quality of the cooling curve and, therefore, the quality of the asphalt. Thus, it is important to have concrete strategies in this section to have a proper base for an adequate analysis.

As mentioned in the previous sections, one of the limitations of the PQi framework is the lack of strategies to collect the data. There is not a clear plan for placing the temperature sensors along and inside the asphalt layers. Information such as thermal conduction coefficients and missing data is currently included by hand. Moreover, some stakeholders are concerned about the efficiency and effectiveness of the results when using different equipment. It is argued that this may hinder the potential benefits of the framework. All of these not only make the analysis difficult but also time-consuming. Therefore, to have a better organization of the modulus of the cooling curve in the PQi framework, some strategies are presented below.

4.2.1. Information to be Collected

The first step is to identify all the essential information that needs to be collected to determine the cooling curve of the asphalt. Therefore, the information was divided into two groups: pre-construction information and on-site construction data. The overview of this division can be seen in Table 3.

The first group includes information that the client must provide, and the information retrieved from hand-written documents needed for the cooling curve calculation. The client provides information about the requirements of the road, including the location of the project, the type of mixture, the desired density, and the length and age of the road, among others. For the cooling curve, some parameters must be needed, such as the thermal conduction coefficient of the materials. This information will be very useful and must be well integrated into the framework not only for the regression of the cooling curve but also for the maintenance and future controls of the built road.

The second group, the on-site construction data, includes all the temperature readings from the asphalt layers, the ambient, density of asphalt after each roller pass, roller model, time of roller pass, and compaction mode. The current setup that the PQi framework uses is good enough for collecting the needed data. Therefore, the same setting and devices will be used to acquire the data. This means the nodes named paved and processing will still be the main collectors.



Table 3. Information to be collected.







Pre-construction data		On-site construction data	
The thickness of the layer	Thermal conductivity	Delivery temperature	Time of roller pass
Binder properties	Specific heat capacity	Paving temperature	Compaction mode
Road geometry	Heat transfer coefficient	Ambient temperature	Sky conditions
Type of mixture	Desire density	Wind speed	Temperature surface
Client information		The temperature of the asphalt layers	Asphalt density after compaction
		Humidity	Pressure

4.2.2. Equipment for data collection

To collect the necessary information, not only human resources are needed but also technology. The PQi framework already has a set of instruments that are used to collect the information. ASPARi, as the framework developer, owns and provides these instruments for each project. Table 4 provides an overview of the equipment used during the asphalt construction.

Table 4. Equipment used for data collection.

Instrument and features	Instrument and features
<p>Thermocouple</p> 	<p>Extech HD200 thermometer</p> 
<p>This is a sensor used to measure temperatures. It consists of two metal wires joined at one end and connected to a thermometer or data logger. It provides measurements over a wide range of temperatures (omega, n.d.).</p>	<p>The data logger saves in an SD card all the temperature readings measured by the thermocouples.</p>

<p>3D stands</p> 	<p>IR-camera</p> 
<p>These stands were designed and printed by the ASPARi crew to stabilize the thermocouples inside the asphalt layers. Its high depends on the asphalt layer thickness to be measured and must be one cm lower.</p>	<p>The infrared camera will be used to measure the surface temperature of the asphalt only at the measurement point.</p>
<p>Compact mini-PC</p> 	<p>LCD screen</p> 
<p>This will be used for collecting, storing, and pre-processing data from the Asphalt node.</p>	<p>This screen displays the user interface of the asphalt node software.</p>
<p>CSB Battery 12 V 30 Ah</p> 	<p>Portable gauge</p> 
<p>The power needed to have all the instruments working properly will be supplied by this battery.</p>	<p>This instrument is used to measure the density of asphalt after each roller pass on the measurement point. This will help to</p>

However, sometimes the client or constructor can also provide or prefers to use their equipment. The use of different technological instruments is a concern for stakeholders since the result may slightly differ from one piece of equipment to another and affect the results. Therefore, to avoid inconveniences related to readings, it is important to ensure that all the instruments that are used during data collection are calibrated before the start of the onsite construction. The objective of calibration is to ensure the accuracy of tests mechanical, electrical, or electronic instruments and to minimize the uncertainty of measurements (Ganesha & Aithal, 2022). According to Ganesha and Aithal (2022), the calibration of a measuring instrument should be done a) according to the specifications of the device's manufacturer, b) after any mechanical or electrical shock, and c) periodically, either annually, quarterly, or monthly.

Therefore, two instruments are the most important to check and have calibrated before each project starts: the thermometer and IR camera. If the ASPARi's instruments are calibrated, these can then be used to compare and calibrate other instruments that the client might want to use. Below, there are some recommended procedures to calibrate each instrument.

- **Extech HD200 thermometer:** According to the manual (Extech, 2007), the procedure should be as follows:
 - a) Plug the thermocouple into the input connector.
 - b) Put the thermocouple in a place with a known, steady temperature, for instance, an ice bath (0°C) or boiling water (100°C).
 - c) Let the readings stabilize.
 - d) Press "Setup" and change the offset until the primary reading matches the calibration temperature.
- **IR camera:** The calibration of an IR thermal camera should be done by the manufacturer. It is suggested to do this at least once a year (Movitherm, 2023) due to the complexity of the process. However, it is possible to regularly check the calibration of the IR camera by using a similar method as the thermometer. Pointing the camera to an ice bath or boiling water, the temperature readings must be 0°C or 100°C, respectively, or very close to these values.

4.2.3. Sensor placement

The asphalt node is one of the most important nodes for the data collection phase. It consists of two main elements. First, a set of thermocouples are placed inside the asphalt layer at different depths and kept in the desired position by using a 3D stand on the base layer (article main). Second, the infrared camera measures the surface temperature of the asphalt. It is crucial to establish certain rules for correct data collection. Thus, some strategies are proposed considering three important aspects for the data collection at this node: a) the location of the measurement points both longitudinally and transversely, b) the time interval of the data collection, and c) the temperature limit to stop the data acquisition.

The location of the asphalt node will be called the measurement point. Each measurement point must have a strategic transversely and longitudinal location for which some aspects must be considered. First, the distance from the edge of the pavement to the node and a representative longitudinal distance to locate each measurement point.

The thermocouples and the stand form a subset of the asphalt node. It must have a suitable location inside the asphalt layer and the pavement area to obtain representative readings but also not affect the asphalt structure and quality since the stand can reduce the strength of the asphalt in that specific area and causes cracks. Therefore, as mentioned before, its height must be one centimeter lower than the actual thickness of the asphalt layer. Moreover, it must be placed as near as allowed to the middle of the pavement area. To avoid damage to

the 3D stand and the asphalt pavement, it is recommended to place the subset at least 30 cm distance from the edge of the pavement (Punurai, 2003).

Furthermore, it is also essential to consider the longitudinal intervals of the measurement points. That is to say, where to place the set of thermocouples and IR camera to obtain representative information that accounts for variations in placement and time throughout the length of the pavement area (Punurai, 2003). Therefore, each measurement point will be located based on the Beschrijvende Plaatsaanduiding Systematiek (Descriptive placeholder system)– BPS-. This is an alternative way to describe places or sectioning roads in the Netherlands. It uses the hectometre signs along the road to determine an exact place on the road. Furthermore, it guides how to collect, and record needed data (Driessen et al., 1994). Thus, based on the BPS, each measurement point will be located 100 m distance from each other. This will help not only to have a precise interval between measurement points but to store that information and use it in future inspections and maintenance of the same road.

All sets of thermocouples will be connected to the thermometer, which is the device that collects the temperature readings and transmits them to the processing node. Additionally, as mentioned before, an infrared camera will be placed, pointing the measurement pointing at 30 cm to collect the surface temperature.

Finally, the acquisition of surface and core temperature will start as soon as the paver passes through the measuring point within a time step of one second until the surface temperature reaches 50 °C. This can be expected to last from 30 to 40 minutes, depending on the asphalt thickness (Bijleveld, 2015).

A minimum of 50 ° C was selected based on the following reasons. First, the analysis of the asphalt cooling behaviour can be assumed to be divided into three phases of compaction: breakdown, intermediate, and finish rolling with the corresponding windows 120 °C, 120 to 90°, and 90 to 60° (Bijleveld, 2015). Additionally, expert operators suggest or predict window compactions till 50 ° for the majority of HMA (Bijleveld, 2015).

Figure 11 shows the setup of all the sensors used in the asphalt node during the data collection.

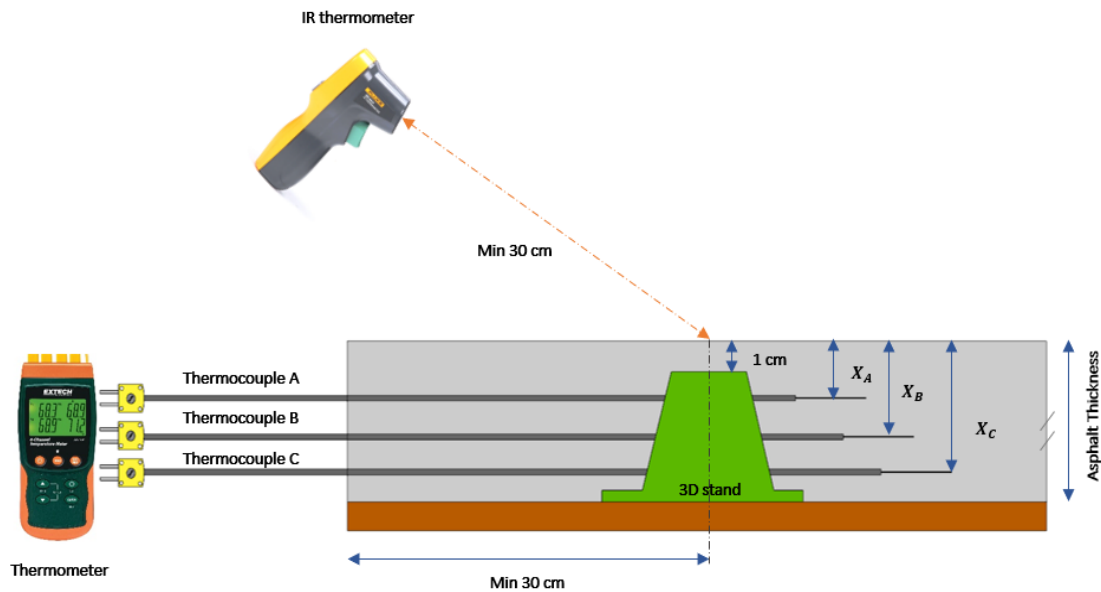


Figure 11. Set up of the asphalt node [adapted from (Makarov et al., 2021)]

4.3. Data preparation pipeline

Data preparation is the largest part of all data management and includes the design of the multidimensional database and the data cleaning methods to ensure the information to be used in the analysis is accurate. These are described in the following paragraphs.

4.3.1. Snowflake scheme

A snowflake scheme is used to represent the relationship between the parameters and the organization of tables for this system. This scheme was based on the current relational database employed by the PQi framework, which can be seen in Annex D, and the parameters that influenced the cooling behaviour found in section 3.1.

In this case, the information will be organized into five dimensions: client, contractor, road, asphalt layers, and measurement points. These groups will serve as answers to the following predefined questions:

- Whom are we working for?
- What are the road features?
- What are the pavement requirements?
- How does the asphalt cool down during the construction?

As seen in Figure 12, almost all dimensions have one or two subdimensions. Additionally, each dimension and subdimension have multiple attributes. For instance, the client dimension has two levels and two attributes: client ID, Name. At the same time, each attribute has an attribute type integer and string, respectively.

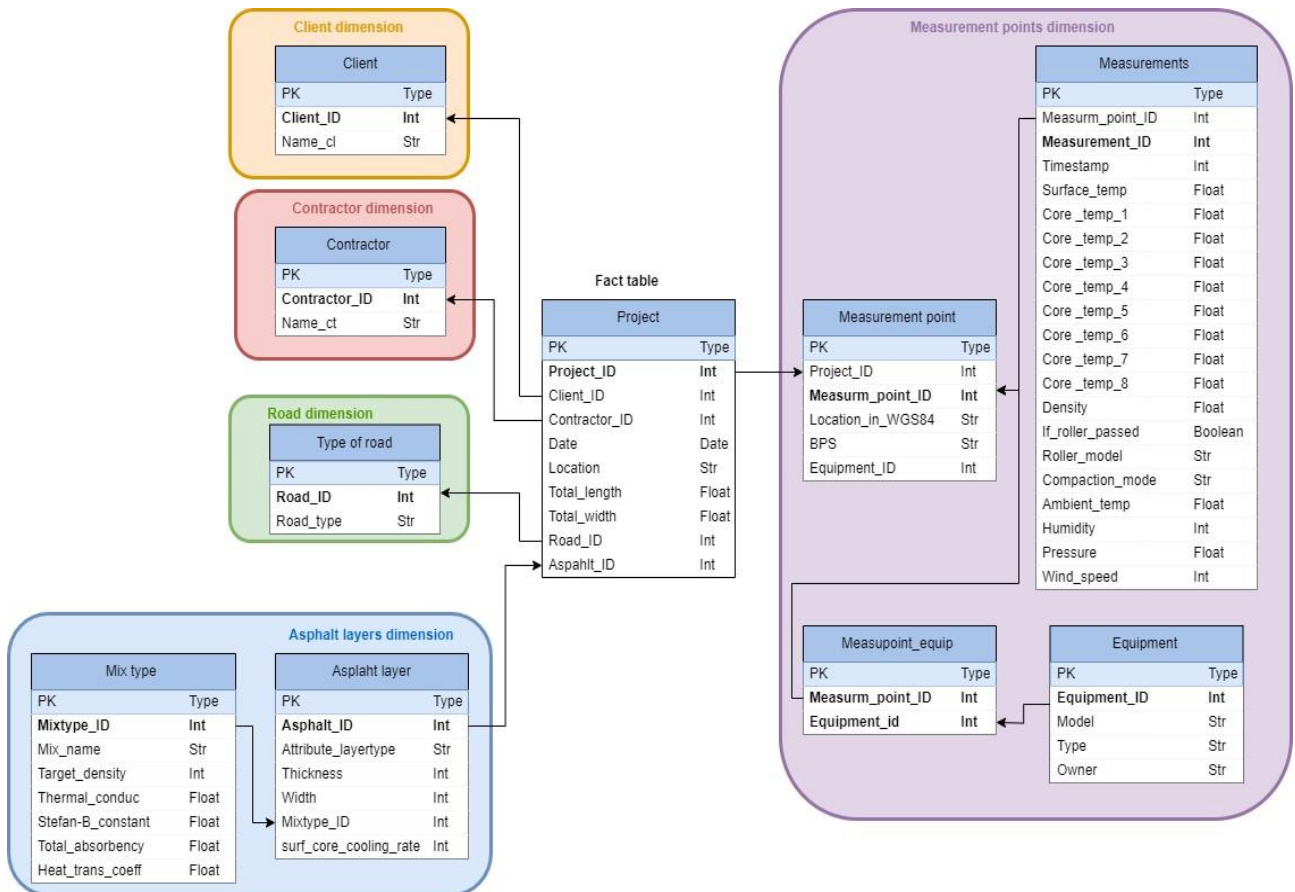


Figure 12. Snowflake scheme of the relational database

When comparing the PQi relational database and the designed snowflake scheme, the following can be mentioned. The PQi framework is not only focused on the cooling curve station but includes the generation of the compaction priority map and compaction contour plot to provide operators with visual aids. Therefore, it consists of more dimensions, such as the paver and roller, which are beyond the scope of this project. Thus, these two were dropped from the snowflake scheme.

Instead of the 'cell' dimension, a 'Measurements' dimension stores the on-site construction data. This data is being employed for the cooling curve estimation. The difference here is that the PQi framework database includes more parameters about the compaction, such as temperature at first/last compaction, number of compaction-passes, remaining compaction passes, time left for compaction, and compaction priority index. All these parameters are used for the visual aids, which are beyond the scope of this project.

4.3.2. ETL process for the system

Extraction

The information collected during the on-site construction will be stored in local devices depending on the equipment that has been used. This raw data will be extracted using Python and its libraries to read the different kinds of files. For packages' specifications, see Annex C. Then, the data will be stored in a data hub or lake without a specific structure. This

data lake will be contained information about the surface temperature captured by the IR camera, the core temperatures from the thermocouples, and the weather station.

To extract the information from the different documents, it is important to understand which type of documents are expected and how the information is structured in those files. Then determine which information is relevant for the purpose of this project.

- **IR camera:** The retrieved file from the IR camera is a text file of n-columns depending on the total time spent during the asphalt construction and 11 columns. From all the information in this file, only the first two columns will be extracted for the database and used to estimate the cooling behaviour of asphalt, which are time and the measured temperature.
- **Thermometer and thermocouples:** The temperatures collected by the thermometer are stored in a text document containing 11 columns. Each column represents the place, date, time, value, or unit. The value columns represent the temperature measured by a thermocouple at a certain time. Since several thermocouples are being used at the same time, the columns "unit" shows which thermocouple measured that temperature. These two columns are relevant for the cooling curve. Thus, this information will be extracted from the document to the database.
- **Weather station:** This is an Excel file that can be retrieved from the Meteorologisch Instituut, Ministerie van Infrastructuur en Watersaat, or from the weather's local station. The document has 38 columns. However, only the columns of time, ambient temperature, pressure, wind speed, and humidity will be useful for this project.

Additional features related to the information already included in some tables and the format of files and folders needed to extract the information using Python can be found in Annex B.

Transformation

As mentioned before, this phase is the most time-consuming. After this phase, it is expected to have more accurate data to generate the asphalt cooling curve later in the process. Thus, the data cleansing for the retrieved information will be focused on dealing with irrelevant data, outliers, and missing data to ensure the predictions of the asphalt cooling curve are accurate.

First, as mentioned in section 4.2.3, the measurement of the temperatures will be done until 50 °C. Thus, temperatures lower than 50° will be removed from the dataset as well as temperatures over 200 °C. Data points lower than 50°C are considered irrelevant for the estimation of the asphalt cooling curve since compaction at these lower temperatures no longer affects the density of asphalt. Thus, excluding these values will ensure the accuracy and appropriateness of the model.

Second, identifying, and correcting outliers with more coherent values will be done graphically using linear regression. According to Lianne and Justin (2020), the outliers are

observed data points considered inaccurate or incorrect values in the data set. Moreover, these points have large errors and are far from the least-squares line (“Linear Regression: outliers,” n.d.). Therefore, to identify these points graphically, two things are used: the best-fit line equation and the data set's standard deviation. The outliers can be easily identified by plotting them using the following rule: the vertical distance of any data point to the corresponding point on the line of best fit are equal to or greater than the standard deviation (“Linear Regression: outliers,” n.d.). Any point outside the rule is considered outliers and removed from the data set. Normally two standard deviations are chosen to check the outliers. However, since the data used in this project is temperature and the variation from second to second should not be significant, one standard deviation was used in the rule to follow the gradual variation.

After the outlier has been removed, a new value is assigned using the mean between the predicted value and the average of the previous ten 10 points minus the average rate of change of the whole dataset. This can be seen in Figure 13 and Figure 14. This is done for the IR camera and thermometer data in Python using the library `sklearn.linear_model: Linear Regression` and `sklearn.preprocessing: Polynomial Features`.

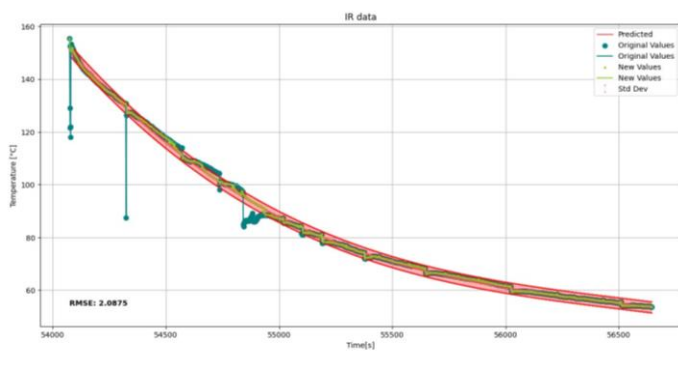


Figure 13. IR camera temperatures plotted including the regression and std dev

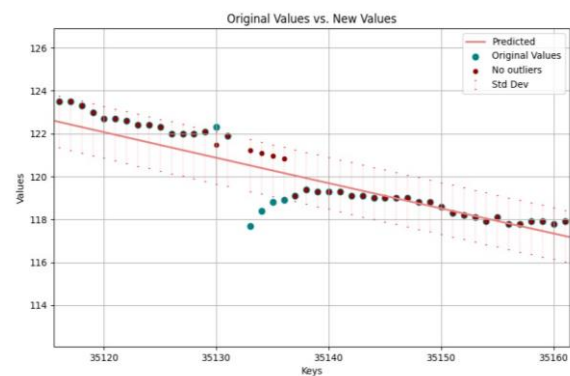


Figure 14. Identifying and correcting outliers

Third, sometimes the dataset provided is incomplete, and some points are empty, as seen in Figure 15. Thus, to handle missing data, the KNN imputation method was used. According to Sahoo & Ghose (2022), KNN imputation deals with the missing points found when working with an incomplete dataset. It selects the k nearest neighbours from a missing point. Then a new value is estimated for the missing datum depending on the data type, mean, and mode of the data. This can be done in Python using the package `sklearn` with the class `KNNImputer`. This class offers imputation using the k -nearest neighbours to replace missing values.

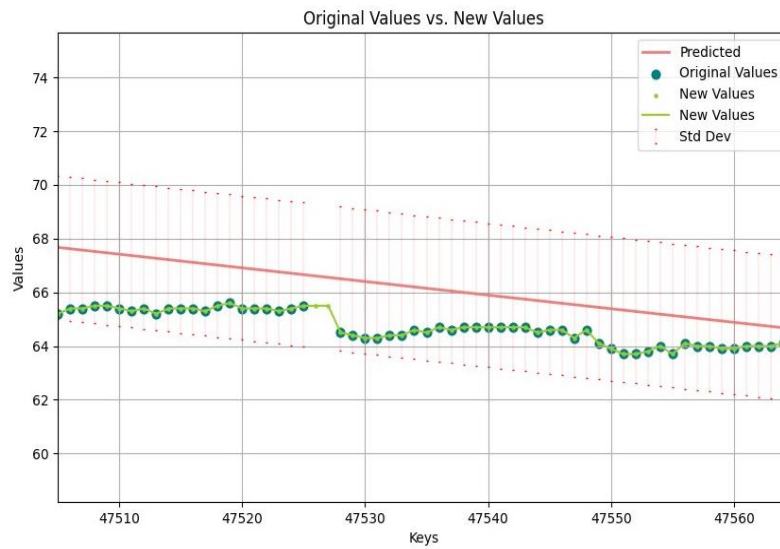
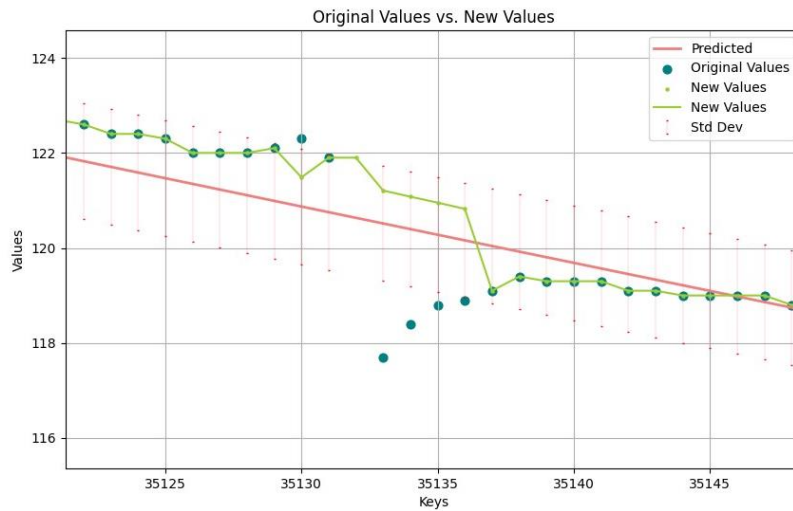


Figure 15. Missing data points

Loading

The database created in pdAdmin 4 from Postgres is called "Asphalt cooling curve." The data that has been already prepared and cleaned is ready to be loaded into this database. It uses SQL language in Python to access the database and load all the necessary information. The local database can be accessed by using the following information:

```
host="localhost",
database="Asphalt cooling curve",
user="postgres",
password="Gamo2207",
port="5432"
```

4.4. Data Analysis Mechanism

After the preparation phase, where the data has been cleaned and loaded into the database, it is ready for analysis. In this case, this is the estimation of the asphalt cooling curve. Due to time constraints, the method selected to perform this duty is an empirical model rather than a theoretical model. Thus, a multivariate polynomial regression will be used to estimate the surface temperature of the asphalt based on the information retrieved during the on-site construction. These are the core temperatures, the ambient temperature, the pressure, the wind speed, and the humidity. This will allow us to estimate the cooling curve based on the observed and measured temperatures. However, a more theoretical approach can be made using additional information that the database contains.

The polynomial regression can be done using the same package in Python as before [sklearn package] (Kaplan, 2022). To select the most suitable regression degree, five different were performed from two to six, and their RMSE values (Table 5) were compared. In the end, a second-degree regression was selected to get the best-fit line per measurement point in a project since the curve was smoother compared to higher degrees.

Table 5. RMSE values of the regressions

Project_ID	MP_ID	RMSE 2	RMSE 3	RMSE 4	RMSE 5	RMSE 6
1	1	0.5264	0.4218	0.5169	0.5402	0.3834
1	2	0.1521	0.1393	0.1285	0.1162	0.1104
1	3	0.7392	0.6360	0.7242	0.5676	0.5317

Thus, to plot the asphalt cooling curve, the program access the data from the database depending on the project id and the measurement point that the user wants. Figure X shows the final plot of the cooling curve that includes the core temperatures of each thermocouple and the surface temperature. Moreover, the *RMSE* and a R^2 for each regression are also calculated to check the performance. More information about these two parameters and how to calculate them can be found in Annex E. An example of the final graph using dummy data can be seen in Figure 16.

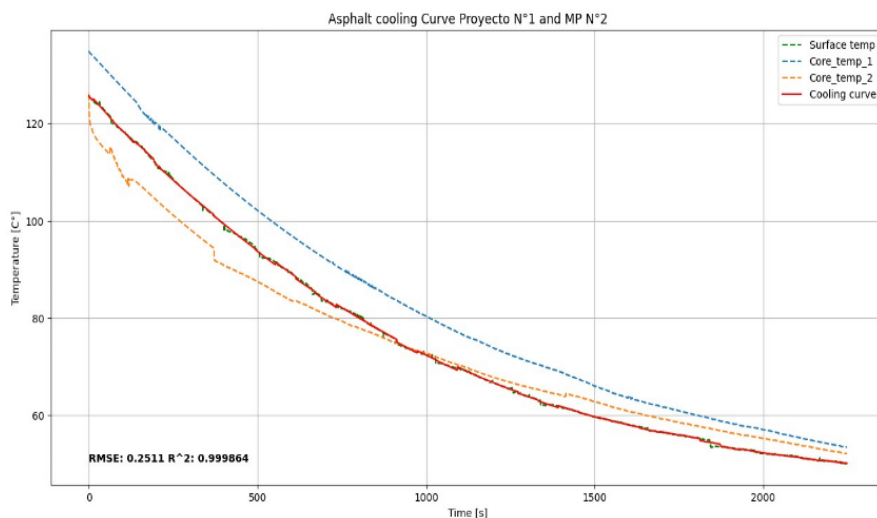


Figure 16. Asphalt cooling curve- dummy data

5. Treatment validation

As mentioned in section 2.4, the validation of the data infrastructure consisted of testing the data infrastructure with real data, then presenting and explaining the results to experts to get feedback. Therefore, this section presents the results of the cooling curve estimation and the opinion given by the experts during the experts' opinion session.

5.1. Case study

The main objective of this case study is to check the performance of the designed data infrastructure by using the data from a real project carried out in the Netherlands in 2022. It is important to mention that for this case study, the proposed data collection strategies could not get tested since the data was already collected beforehand and handed in. However, it is assumed that the strategies can benefit the systematization of the collection process.

This project had four measurement points where the data was collected. Therefore, four cooling curves are expected. The following figures represent the data from the first MP. The graphs of the rest of the MPs can be found in annex G. Thus, Figure 17 shows the raw data from the IR camera. Here the outliers and missing data points can be seen. On the other hand, the data after the transformation phase of the designed data infrastructure is displayed in Figure 18. It can be seen that the data follows a smoother trend without significant outliers and is complete within the time series. The same process was applied to the three other MPs, as Figure 19 shows the after transformation, and to the information from the thermometer.

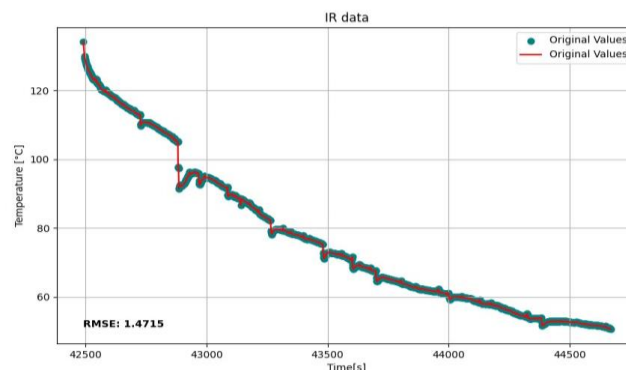


Figure 17. Raw IR camera data

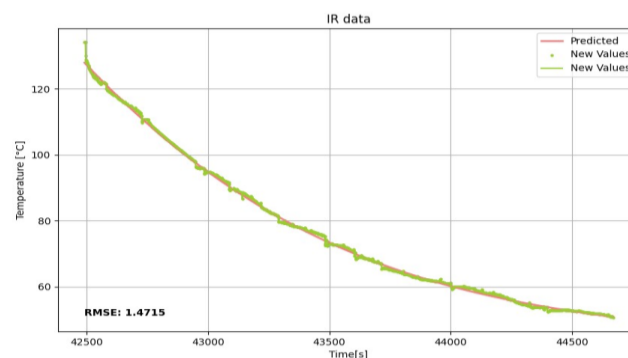


Figure 18. After transformation IR camera data

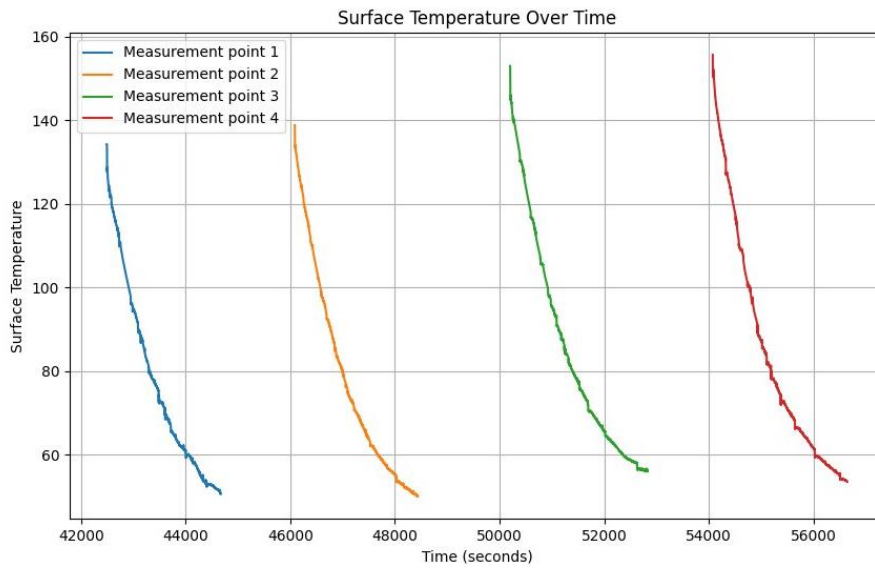


Figure 19. After transformation surface temperature

The main plot and the only one that can be seen at the end of the data infrastructure is the asphalt cooling curve, which is shown in Figure 20. Two more features were included in this graph the regression equation and the time intervals in minutes. The values for RMSE and R2 are 0.38 (low) and 0.99 (high), respectively, which is the desired situation. This means that the predictions of the regression have a strong relationship with the response variables of the model.

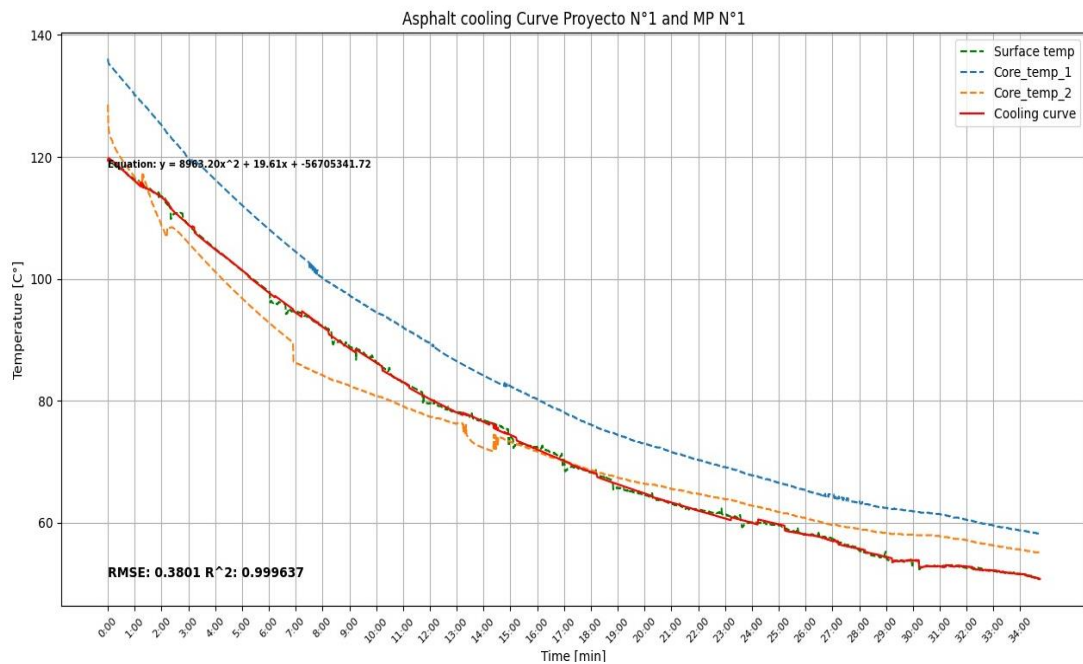


Figure 20. Asphalt colling curve MP 1

5.2. Expert Opinion

As mentioned in section 2.4.2, the expert opinion session allowed the experts to evidence and evaluate the performance of the designed data infrastructure. At the same time, the designer gains insights, feedback, and guidance for further improvements and developments.

The session was done with three stakeholders' representatives. Figure 21 displays the average results of the quantitative evaluation using the pedigree matrix per criteria. The individual evaluation is in Table 6, annex F. The general grade for the entire data infrastructure is 3.3 out of 4. The stronger point was the theoretical basis. The expert agreed with the theory used to build the model, as well as the methods for both the data preparation and the estimation of the cooling curve, since these are methods that are commonly used. The lowest grade of 3 is for the completeness and plausibility of the model. The experts argued that even though the model does capture the main parameters that play a role in the asphalt cooling behaviour, the necessities of the client can vary from project to project. This will require more input parameters to meet those requirements. It was also mentioned that additional features, such as the density progression, can be added to have a complete analysis of the construction process and to observe the relationship of the cooling behaviour, compaction, and density.

The answers to the questionnaire reflected a good overall impression of the design. This more automated system reduces the bias induced by human interference. The modularity of the system is seen as a strength since this gives more freedom to adjust or make changes in the model if needed. Additionally, the model was considered reliable due to the employed methods, and by comparing the plots with similar cooling curves the experts have seen before.

The challenges that could be faced when implementing the system in the real world are incorporating the measured cooling curve during the construction as an input for the guidance of the roller operators. Therefore, it was suggested to bring focus on the prediction of the cooling curve before the construction start and consider the clients' and contractors' requirements. Also, deal with the biased operators who optimize their work right on the MP when someone is monitoring the work, but performance declines afterward. It was also pointed out that to understand the system, a certain level of programming knowledge will be needed. Thus, to make it more user-friendly, a dashboard or a manual was suggested. These would enhance the usability and navigation of the system for external UT members and non-experts.

ACC data infrastructure evaluation

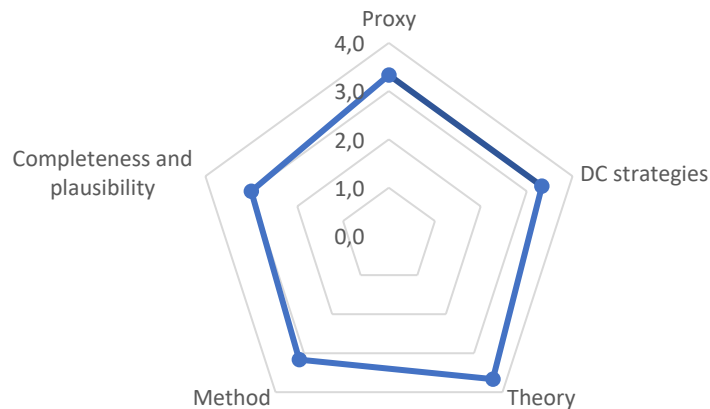


Figure 21. Pedigree matrix results

6. Discussion

- Results of the system validation

The asphalt cooling curve station of the PQi was studied during this design project. First, to identify its weaknesses and then to propose and develop solutions to transform those weaknesses into strengths. In this context, the main contribution of the research was the development more systematic and automated data pipeline that considers all the steps of data management. The data collection, preparation, visualization, and use of that data.

The system was well-received by the experts, demonstrating a strong degree of acceptance. As mentioned in section 5.2, the overall given grade of the system was 3.3. Experts pointed out that the strengths of the system are the modularity of the scripts, the theoretical basis, and the methods used to calculate the asphalt cooling curve. On the other hand, the weaknesses of the system lay in two things that were also mentioned by the experts. a) Prediction of the cooling behaviour before the actual construction is unable, and b) The lack of a dashboard or a more user-friendly interface that facilitates the navigation through the system. These two functions were not implemented in the system due to time constraints and complexity.

The final data infrastructure was designed based on the requirements stated in section 4.1. They included, first, the integration of strategies during the on-site data collection. Even though these strategies could not be verified and validated during the case study, the experts argued that the strategies are consistent with the stakeholder needs and easy to implement.

Second, an automated data preparation pipeline that not only automates the detection and imputation of outliers and missing data but also ensures automatic data integration. Therefore, to organize all the needed parameters, a multidimensional relational database was done based on the relational database proposed by Makarov et al. (2021). Several

modifications were needed since its focus goes beyond the scope of the project. Thus, only parameters relevant to the cooling curve were considered. Moreover, some challenges were faced during this task, such as confusion in choosing software for building the database and difficulties finding the right approach for data cleansing, which consume more time than expected. However, experts positively accepted the methods used for both outliers and missing data detection. The use of both open-source software, python and PostgreSQL, was also considered a favourable feature that avoided license issues.

The last requirement was the automatic data analysis pipeline. The multivariable polynomial regression was approved during the expert opinion session. They based the answer on the low values for the RMSE-around 0.2 to 0.6 - and the high values of R². This could be changed when making a better prediction of the curve before the actual asphalt construction. Additionally, the prediction of the cooling curve before the actual construction was not included due to time limitations. However, the asphalt cooling curve estimated by the system can be used for the post-analysis of the pavement and to train other models for predictions.

- Implication for the framework

This design project was part of the methodology followed by ASPARi, which is called the Process Quality improvement -PQi-. This cyclical method focuses on defining, measuring, analysing, getting feedback, improving, and enriching the employed methods by increasing the knowledge about road construction and involving students from different levels in the subject. This is done to enhance technologies and processes that improve the quality of the roads (ASPARi, n.d.). In this context, the ASPARi research network developed the PQi framework, which is a tool that wants to improve construction quality by increasing the effectiveness of compaction—a critical determinant of the quality of asphalt during construction and afterward. Therefore, the aim is to increase the use of technologies and decrease the (biased) human intervention to ensure the quality of compaction and, therefore, the quality of pavements.

To measure the quality of the compaction process, the PQi framework employs the effective compaction rate - ECR - index. It shows how well the asphalt layer was compacted in relation to the percentage of the layer that was done so within the proper temperature window (Makarov et al., 2021). That is to say, and the ECR considers the sections of the pavement that had the desired roller passes done within the compaction window.

As mentioned in section 3.1., the temperature window is calculated and displayed in the cooling curve. Hence, the importance of this curve in the overall quality of asphalt and the pre and post-analysis of the pavement. In this context, the main contribution of this project lies in the development of the automated data infrastructure for the cooling curve station of the PQi framework. It enhances the accurate estimation of the cooling curve from its structural foundation until the end of the data life cycle. This infrastructure ensures the availability of quality data to work with, free of misleading data points that can affect the overall result.

Employing a systematic data collection approach facilitated adherence to an established or standardized procedure for data acquisition, which constituted the initial step to obtaining qualitative data. The transformation phase employs widely adopted techniques to ensure data integrity by identifying, eliminating, and replacing outliers and by handling missing values. Subsequently, the processed data is employed in plotting the cooling curve. Although predictions are not explicitly incorporated into the model, the gathered data and the asphalt cooling curve bear considerable significance for subsequent analysis of the construction and model training.

7. Conclusions and future work

The objective of this project was to develop an automated data infrastructure that facilitates the collection, storage, integration, processing, and analysis of the asphalt cooling process and data. Thus, by considering the limitations present in the PQi framework, such as lack of data management and processing. The main contribution of this research was the development of a more suitable data pipeline that focuses on providing a solid and improved foundation for the asphalt cooling curve station of the PQi framework. To develop this data infrastructure, the method of engineering design cycle was adopted. It helps to identify the relationship between the activities and the input-outputs for each phase: problem investigation, treatment design, and treatment validation.

Therefore, the project started by studying and examining the theoretical framework and the social context, which served as the foundation of the data infrastructure. Having the stakeholders' needs, the requirements of the system were defined, and the data structure built. First, a standardized data collection process was suggested to ensure the same process was followed during each pavement project. It included establishing which parameters are required, the equipment to use, the distribution of the MPs, and sensor placement.

The data preparation required a substantial investment of time and resources. But for this task, multidimensional data management and the ETL were key concepts to understand. The methods used to deal with outliers and missing data points were polynomial regression and KNN imputation, respectively. This process improved the quality of data to be uploaded to the built database. Then, the estimation of the cooling curve was done using a conceptual model. This is a multivariable polynomial regression that considers the core temperatures, ambient temperature, humidity, pressure, and wind speed.

In the end, the system was validated using real data and elicited by experienced people. The experts gave the system a positive review, indicating a good level of acceptance. Some positive and negative features of the system were pointed out. The modularity of the system's components, its theoretical underpinnings, and its methodologies for calculating the asphalt cooling curve, according to experts, are its strong points. On the other hand, the identified system's shortcomings were the lack of a dashboard or user-friendly interface that makes the use of the system easier for experts and non-experts. Also, the lack of cooling curve prediction before the start of the construction phase. Due to complexity and time constraints, these two functions were not included in the system.

Overall, the designed data infrastructure fulfilled almost all the requirements in terms of automatization of the data preparation pipeline. Nonetheless, it encounters challenges with preconstruction predictions of the asphalt cooling curve for contractors. However, the core objective of the project, which was improving the foundations of the asphalt cooling curve station and data management, was achieved. An organized database and better data preparation pipeline will result in a more accurate cooling curve, optimize the compaction process, and ensure the quality of asphalt during and after the construction phase. These

advantages would be reflected in the ECR index used by the PQi framework. Moreover, the system can be combined with new projects focused only on predictions and support the analysis and study of cooling behaviour in specific asphalt mixtures.

7.1 Future work

The subsequent steps for this project involve the incorporation of the asphalt density progression during the construction to evidence and establish the correlation between density and temperature. Moreover, integrating the design into the PQi framework in combination with the temperature and compaction contour plot and the compaction priority map. Additionally, incorporating additional predictable features for the asphalt cooling curve before the construction process would be very useful for contractors. Since it will enable them to offer better compaction guidance to operators and ensure superior asphalt quality. The prediction will have to be based on weather prediction and to consider the asphalt thickness and type of mixture.

It is also recommended to expand further the validation of the system. Since the validation of the design relied only on data from one specific project, it is advised to conduct more tests and comparisons with similar tools or studies. Therefore, the model's robustness will increase by putting the data infrastructure under a more rigorous validation and verification process that checks the accuracy of the underlying theories, inputs, and outputs.

8. References

- Arbeider, C. (2017). *Planning the asphalt paving and compaction process "The alignment between paver output, roller capacity and available time for compaction."* Enschede.
- Ayuquina, K., (2022). *Predicting asphalt temperature for construction A contribution to the analysis of the cooling rate of asphalt during the construction process.* [Research thesis, University of Twente]
- ASPARI. (n.d.). Over ASPARI. Retrieved April 11, 2023, from <https://en.aspari.nl/about>
- Bijleveld, F. (2015) *"Professionalising the asphalt construction process. Aligning information technologies, operators' knowledge and laboratory practices"* [Research thesis, University of Twente]. University of Twente Research Information. <https://research.utwente.nl/en/publications/professionalising-the-asphalt-construction-process-aligning-infor>
- Chadbourn, B. A., Newcomb, D. E., Voller, V. R., Desombre, R. A., Luoma, J. A., & Timm, D. H. (1998). "An asphalt paving tool for adverse conditions." Minnesota Dept. of Transportation Final Report MN/RC-1998, 18.
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development Discussions*, 7(1), 1525–1534. <https://doi.org/10.5194/gmdd-7-1525-2014>
- Coetzee, S. et al. (2020) "Stakeholder analysis of the governance framework of a national SDI dataset—whose needs are met in the buildings and address register of the Netherlands?". *International Journal of Digital Earth*, 13(3), pp. 355-373. Available at: <https://doi.org/10.1080/17538947.2018.1520930>
- van Dee, R. (1999) *"Modelling of the compaction of asphalt layers"* [Research thesis, University of Delf] Available at: <https://repository.tudelft.nl/islandora/object/uuid%3A43b529c8-edba-447c-aa35-feeca89625a6>
- Driessen, J., et al. (1994). *Beschrijvende Plaatsaanduiding Systematiek*. Ministerie van Verkeer en Waterstaat
- Extech, (2007). *User's Guide Differential Thermometer Datalogger Model HD200 Model HD200 Version 2.0.* (2007). https://www.instrumart.com/assets/HD200_UM.pdf
- Freeman, R. E. 1984. *Strategic Management: A Stakeholder Approach*. New York: Cambridge University Press.
- Ganesha, H. R., & Aithal, P. S. (2022). Choosing an appropriate data collection instrument and checking for the calibration, validity, and reliability of the data collection instrument before

collecting the data during the ph.D.. program in India. *International Journal of Management, Technology, and Social Sciences*, 497–513. <https://doi.org/10.47992/ijmts.2581.6012.0235>

Introduction to SQL (n.d.) *SQL Introduction*. Available at: https://www.w3schools.com/sql/sql_intro.asp

Jensen, C.S., Pedersen, T.B. and Thomsen, C. (2010) *Multidimensional databases and Data Warehousing*. U. st.: Morgan & Claypool Publishers.

Juna, S. (2020). *The problem of overfitting in the prediction of the cooling rate of asphalt mixes within the ASPARiCool tool's MLP algorithm*. [Bachelor thesis]. University of Twente. Available at: <http://essay.utwente.nl/83499/1/Juma-Shaffie.pdf>

Kang, Y. (2010) “*The function tree analysis for new product development & its applications*”. Available at: https://r1.nubex.ru/s828-c8b/f1652_ca/Thesiscomplete_revised%200625_007.pdf

Kaplan, D. (2022). *Multivariate Polynomial Regression Python (Full Code)*. Enjoymachinelearning.com. Available at: <https://enjoymachinelearning.com/blog/multivariate-polynomial-regression-python/>

van Keulen, M. and Ahmed, F. (2021) “Data Science- Topic DPV -Data preparation and visualization” [PowerPoint presentation, University of Twente]

Kimball, R., & Caserta, J. (2004). *The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data*. Wiley. <https://ia800206.us.archive.org/15/items/2004TheDataWarehouseETLToolkitRalphKimball/2004%20-%20The%20Data%20Warehouse%20ETL%20Toolkit%20%28Ralph%20Kimball%29.pdf>

Lianne and Justin (2020) “Data Cleaning in Python: the Ultimate Guide,” *Medium*. Medium. Available at: <https://towardsdatascience.com/data-cleaning-in-python-the-ultimate-guide-2020-c63b88bf0a0d> (Accessed: April 22, 2023).

Linear Regression: Outliers | Saylor Academy. (2023). Saylor Academy. Available at: <https://learn.saylor.org/mod/book/view.php?id=55086&chapterid=40788>

Makarov, D. et al., (2021). A framework for real-time compaction guidance system based on compaction priority mapping. *Automation on Construction*, 129(1), 103818. <https://doi.org/10.1016/j.autcon.2021.103818>

Marsden, E. (n.d.). *Regression analysis using Python*. <https://risk-engineering.org/static/PDF/slides-linear-regression.pdf>

Movitherm, (2023). *Performing a Thermal Camera Calibration*. MoviTHERM. <https://movitherm.com/knowledgebase/thermal-camera-calibration/>

- pgAdmin. (2023). *pgAdmin 4 — pgAdmin 4 7.4 documentation*. Pgadmin.org. <https://www.pgadmin.org/docs/pgadmin4/development/index.html#:~:text=pgAdmin%20is%20the%20leading%20Open,and%20use%20of%20database%20objects>.
- Punurai, S. (2003) Optimization of very early strength concrete mixes using maturity method [Master's thesis]. New Jersey Institute of Technology. Available at: <http://archives.njit.edu/vol01/etd/2000s/2003/njit-etd2003-055/njit-etd2003-055.pdf>
- Sahoo, A., & Ghose, D. K. (2022). *Imputation of missing precipitation data using KNN, SOM, RF, and FNN*. 26(12), 5919–5936. <https://doi.org/10.1007/s00500-022-07029-4>
- Shaker El-Sappagh, Hendawi, A. M., & Ali El Bastawissy. (2011). *A proposed model for data warehouse ETL processes*. 23(2), 91–104. <https://doi.org/10.1016/j.jksuci.2011.05.005>
- Shravankumar Hiregoudar. (2020, August 4). *Ways to Evaluate Regression Models - Towards Data Science*. Medium; Towards Data Science. <https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70>
- van der Sluijs, J., Craye, M., Funtowicz, S., Ravetz, J., & Risbey, J. (2004). Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: the NUSAP system. *Risk Analysis: An Official Publication of the Society for Risk Analysis*. doi:doi.org/10.1111/j.1539-6924.2005.00604.x
- Vargas, A. and Timm, D. (2011). "Validation of Cooling Curves Prediction Model for Nonconventional Asphalt Concrete Mixtures," *Transportation Research Record* 2228:1, pp.111-119 Available at: <https://journals.sagepub.com/doi/epdf/10.3141/2228-13>
- Vasenev, A., Bijleveld, F. and Dorée, A. (2012) "Eurasphalt & Eurobitume Congress," in *A real-time system for prediction cooling within the asphalt layer to support rolling operations*. 5th ed. Turkey, Istanbul: ResearchGate, pp. 1–7.
- Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). Conceptual modeling for ETL processes. *Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP - DOLAP '02*. <https://doi.org/10.1145/583890.583893>
- Wieringa, R. (2014) "Empirical research methods for technology validation: Scaling up to practice," *Journal of Systems and Software*, 95, pp. 19–31. Available at: <https://doi.org/10.1016/j.jss.2013.11.1097>.
- Zhang, Y. and Thorburn, P. (2022) "Handling missing data in near real-time environmental monitoring: A System and a review of selected methods," *Future Generation Computer Systems*, 128, pp. 63–72. Available at: <https://doi.org/10.1016/j.future.2021.09.033>.

Annex

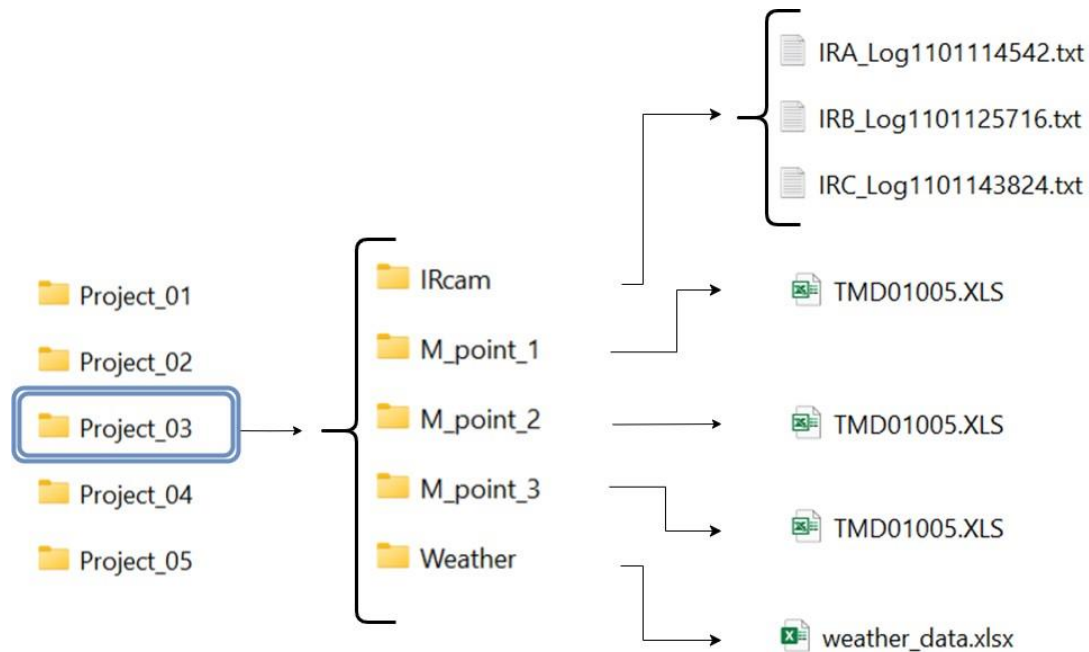
A. ETL framework and process

- **Extraction:** This is the first step in the ETL process. It consists of extracting the data from different sources. These sources might be relational and non-relational databases, flat files, XML data sets, or spreadsheets (Shaker et al., 2011). Thus, to have an efficient data extraction, it is important to identify and understand the features of each source to apply the right approach to manage that specific kind of data (Kimball & Caserta, 2004). According to Shaker et al. (2011), there are two subphases in the extraction process the initial extraction and the incremental extraction. The first is filling the DW, just after its construction, with all raw data from the sources. The second is also known as changed data capture (CDC). It updates the DW with the modified and new data that has been extracted from the sources since the last extraction.
- **Transformation:** This is the second step that consists of performing cleaning and conforming to the entering data. The main goal is to add value to that data by guaranteeing accuracy. That is to say that the ETL system ensures the data is correct, unambiguous, consistent, and complete (Kimball & Caserta, 2004).
- **Loading:** loading is the ETL's final step. It consists of moving the extracted and transformed data to the fact and dimension tables located in the actual DW. At this point, the users is able to access high quality data according to carry out analyses or any other task.

B. Extra features of the system

- Folder and file format to be read by the program.

For the program to be able to read the information from the different devices, it is important that the names of the folders and files are correctly named following the format below shown. Otherwise, the program will not be able to open the files and read the data. Thus, each project will have a project file that specifies its name. Inside the project file several folders can be found named: IRcam, M_point_X (X addresses the measurement point that the file corresponds to), and Weather.



- Information available to select from the tables in the database.

Client_ID [PK] bigint	Name_cl text	Contractor_ID [PK] bigint	Name_ct character varying	Road_ID [PK] text	Road_type text
1	Gemeente Almere	1	BAM	RW	Rijksweg
2	Gemeente Amsterdam	2	Ballast Nedam	PW	Provinciale weg
3	Gemeente Enschede	3	Boskalis	GW	Gemeenteweg
4	Gemeente Oude IJsselstreek	4	Bosklais	WW	Waterchapsweg
5	Gemeente Tiel	5	Dura Vermeer	PA	Particuliere weg
6	Prov. Utrecht	6	Heijmans		
7	Provincie Gelderland	7	KWS		
8	Provincie Noord Holland	8	Lansink		
9	Provincie Overijssel	9	MNO		
10	RWS	10	Mourik		
11	Rijkswaterstaat	11	NTP/Wegenbouw Lansink B.V.		
12	Universiteit Twente	12	Ooms		
		13	REEF		
		14	Roelofs		
		15	Strabag		
		16	TWW		
		17	Van Gelder		
		18	Wegenbouw Lansink B.V.		

Mixtype_ID [PK] bigint	Mix_name text	Stefan-B.constant double precision	Total_absorbency double precision	Heat_trans_coeff double precision
1	AC8	[null]	[null]	[null]
2	AC11	[null]	[null]	[null]
3	AC16	[null]	[null]	[null]
4	AC22	[null]	[null]	[null]
5	PA4	[null]	[null]	[null]
6	PA5	[null]	[null]	[null]
7	PA8	[null]	[null]	[null]
8	PA11	[null]	[null]	[null]
9	PA16	[null]	[null]	[null]
10	SMA-NL	[null]	[null]	[null]
11	SMA-NL 11B	[null]	[null]	[null]
12	SMA-NL 8G	[null]	[null]	[null]
13	SMA-NL 5	[null]	[null]	[null]
14	SMA 11A	[null]	[null]	[null]
15	ECOSTAB 214	[null]	[null]	[null]
16	TS	[null]	[null]	[null]
17	EME22	[null]	[null]	[null]
18	LEAB16	[null]	[null]	[null]
19	2I ZOAB	[null]	[null]	[null]

C. Python libraries and scrips description

- Packages or libraries used during coding.

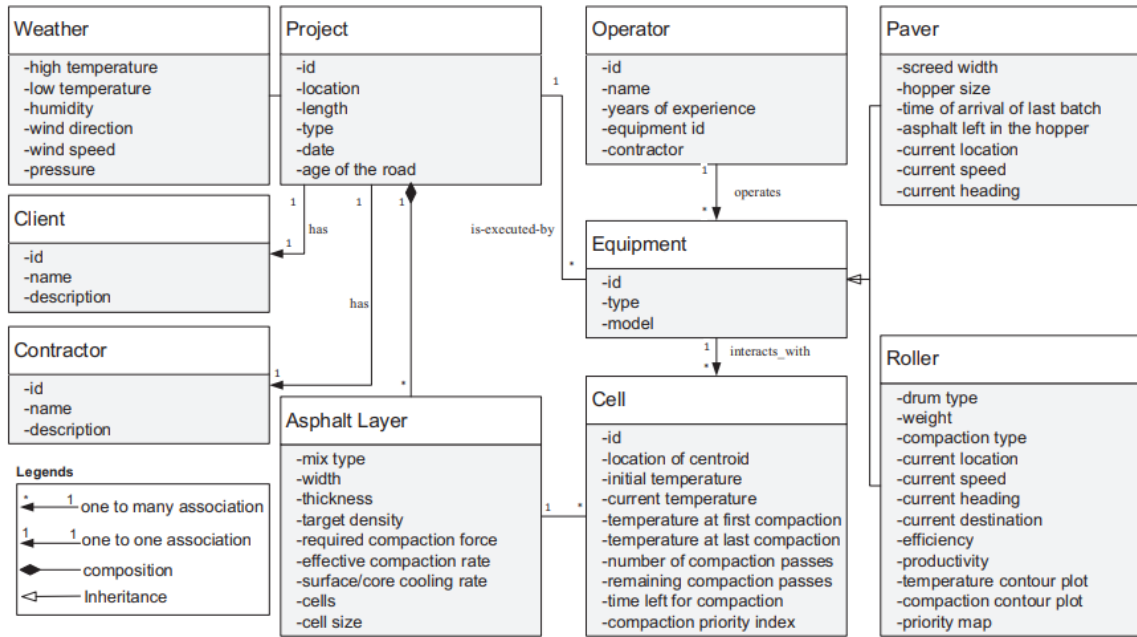
Package name	Description
pandas	pd: used to analyse, clean, explore and manipulate data sets.
Matplotlib	.pyplot: to create visualizations in Python.
NumPy	Np: used for working with arrays and matrices.
sklearn	This helps to implement machine learning models and statistical modelling.
os	Interacting with the operating systems
psycopg2	Database adapter

- Scrips and their use.

Script name	Description
strat_project_inputs.py	This script allows us to start a new project, the user may enter the details of the project such as date, client, contractor, location, and type of road. After running this script all the inputs will be already saved in the database → Table Project
inputs_m_points.py	This script allows us to enter the number of measurement points (MP) that the project will have, including the BPS, and some additional details such as each asphalt layer. Here as well, the data will be automatically saved in the database.
IRdata.py	This script reads the IR camera data from the project folder. Then clean all the data and gives as output a list that contains as many dictionaries as MP are in the project. [{time:temp},{time:temp},...]
THMdata.py	This script reads the data from each of the folders of the MPs. Then, it prepares and cleans all these data. The output is also a list that contains as many dictionaries as MPs. Each dictionary contains the time as a key and a value list with all the core temperatures depending on the number of thermocouples. The format is [{time: [th1,th2,thx]},{time:[th1,th2,thx]}, ...]
Weather.py	The information from the weather station is read and prepared here. The output is a dictionary that contains the time as a key and a list of values that includes, the ambient temperature, humidity, pressure, and wind speed of almost the whole day per minute. {time:[ambt_temp, hum, press, wind_s]}
Jointtemps.py	In this script, all the information from their camera, thermometers, and weather station is put together in a dictionary by considering the time of the day.
Temps_to_DB.py	This script uploads all the data that was put together in the previous file into the database.

Cooling2.py	This script was done to perform the estimation of the cooling curve. It first accesses the data from the data based then calculates the regression and plots everting depending on which project id and which MP id the user inputs.
Import_file.py	This script was used to read the information from clients and contractors from the Excel files provided by ASPARI.
upload_script.py	This was used to upload the information from clients and contractors into the data based. Moreover, it contains information about the road types, equipment, and type of mixture that was also stored in the database. Any additional information needs to be added to these tables can be done using this file.
ACC	This file allows us to run the following files together: strat_project_inputs.py inputs_m_points.py IRdata.py THMdata.py Weather.py Jointtemps.py Temps_to_DB.py

D. PQi relational database



E. RMSE and R2

The RMSE -the root mean squared error- is a standard statistical metric that measures a model's performance using the equation below (Chai & Draxler, 2014). This means that the lower the value obtained, the better the regression predictions.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

Where:

n = number of observations (each observation is one second)

\hat{y}_i = Predicted values

y_i = Observed values

The R^2 – the coefficient of determination- was used to show the proportion of variance of the parameters used in the regression (Shravankumar Hiregoudar, 2020). The value is calculated using sklearn.metrics which employs the following formula:

$$R^2 = 1 - \frac{SS_{regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$



QUESTIONARY



Based on your expertise and experience, what are your overall impressions of the data infrastructure that has been developed for estimating and generating asphalt cooling curve?



As ASPARi member/participant to what extent do you agree (scale from 0 to 5) that this data infrastructure aligns with the necessities and practices carried out by the PQi framework? Are there any major gaps or areas for improvement that you would recommend?



In terms of data collection, processing, modelling, or system integration, can you identify any potential limitations or weaknesses in our data infrastructure?



Do you think the system is easy to use and understand by experts and non-experts?



QUESTIONARY



Are there any data sources/types or methods used that you believe may introduce biases or inaccuracies in the estimation of the cooling curve?



Do you think any challenges or barriers could arise when applying it in real-world scenarios? How can these challenges be addressed?



Can you provide any insights on how the infrastructure's results compare to your own expert judgments or other established methods in terms of accuracy and reliability?

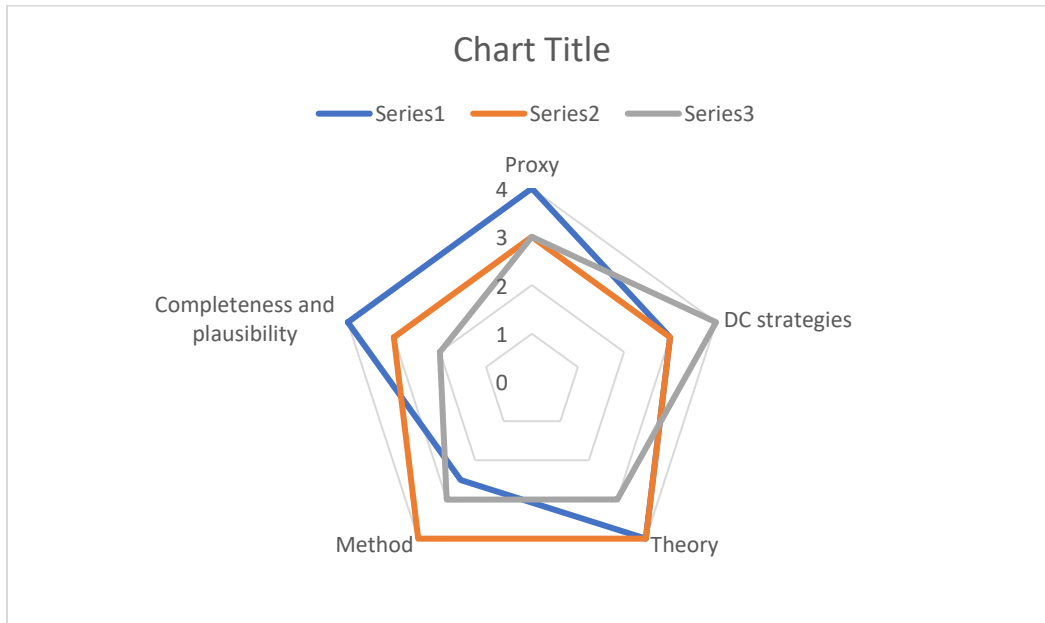


In your opinion, what are the next steps or areas of focus for further improving and evolving the data infrastructure for estimating and generating asphalt cooling curves?

Code	Proxy (indicator of what we want to estimate)	Data collection strategies	Theoretical basis	Method	Completeness and plausibility
4	Excellent	Excellent	Well established theory	Best available practice	Complete and very plausible data pipeline
3	Good	Good	Accepted theory partial in nature	Reliable method commonly accepted	Complete and plausible data pipeline
2	Fair	Fair	Partial theory limited consensus on reliability	Acceptable method limited consensus on reliability	Incomplete but acceptable data pipeline
1	Poor	Poor	Weak theory or concepts, controversial empirical support	Preliminary methods unknown reliability	Incomplete and hardly plausible data pipeline
0	Unacceptable	Unacceptable	No theory or concepts	No discernible rigour	Incomplete and speculative data pipeline

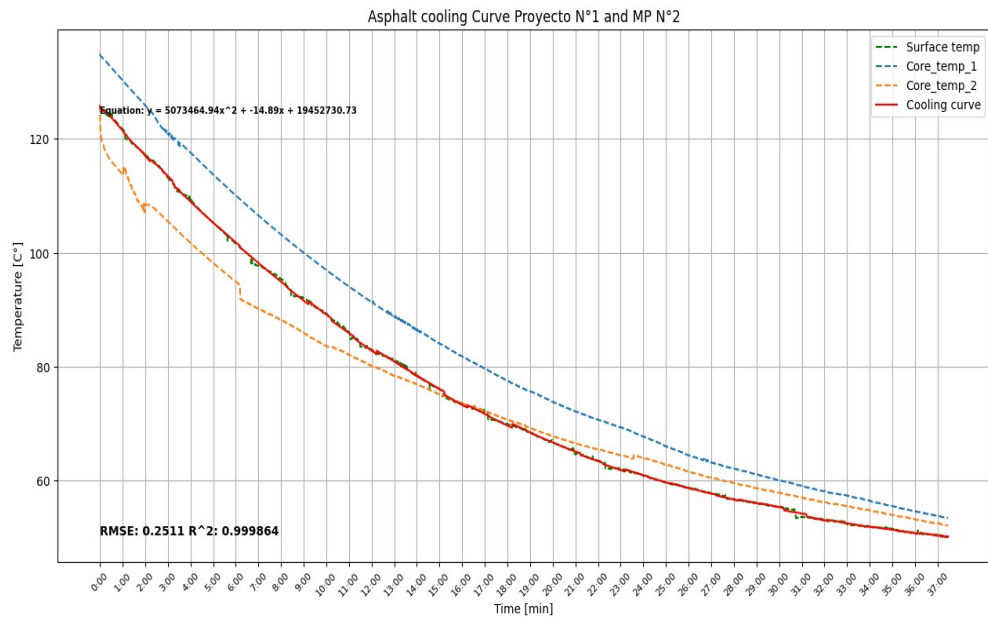
Table 6. Pedigree scores

	Proxy	Data collection strategies	Theory	Method	Completeness and plausibility
Stk. 1	4	3	4	2,5	4
Stk. 2	3	3	4	4	3
Stk. 3	3	4	3	3	2
Final Score	3,3	3,3	3,7	3,2	3,0

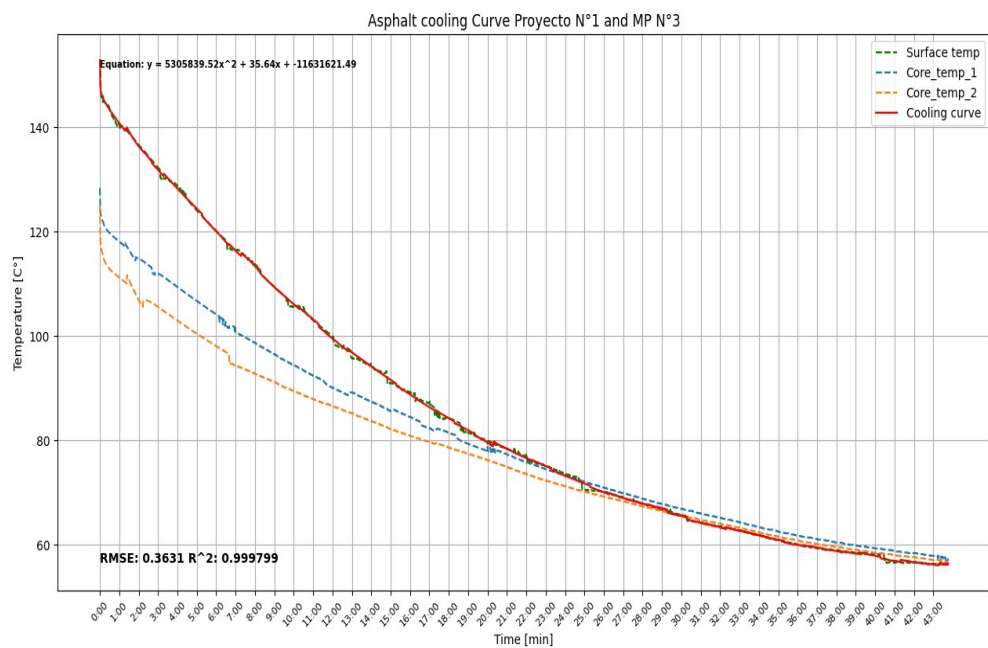


G. Case Study

- Measurement point 2



- Measurement point 3



- Measurement point 4

