Energy Research Programme (Energieforschungsprogramm)

Publishable Final Report

Programme Steering Committee:
Climate and Energy Fund (Klima- und Energiefonds)

Programme Administration:
Austrian Research Promotion Agency
(Österreichische Forschungsförderungsgesellschaft mbH (FFG))

Final Report
created 2022-11-30

EnableDigitalDH

Data science for better data quality: Missing link for a successful data-driven digitalization of district heating

FFG project number: 881 128

Citation (APA):
Hamilton-Jones, M., Falay, B., Tschopp, D., Leppin, L., Toller, M., Macher, C., Pimas, O. (2022). EnableDigitalDH. Data science for better data quality – Missing link for a successful data-driven digitalization of district heating. Final Report FFG Project 881 128. Gleisdorf: AEE INTEC.

Call	Energy Research Programme – 6th Submission
Project start	2021-03-01
Project end	2022-08-31
Project duration	18 months
Project coordinator	AEE INTEC
Contact person	Marnoch Hamilton-Jones, MEng
Address	Feldgasse 19, A-8200 Gleisdorf
Phone	+43 (0) 3112-5886-0, ext. 226
Fax	+43 (0) 3112-5886-18
E-mail	m.hamilton-jones@aee.at
Website	www.aee-intec.at

EnableDigitalDH

Data science for better data quality: Missing link for a successful data-driven digitalization of district heating

Authors:
Marnoch Hamilton-Jones, MEng (AEE INTEC)
Basak Falay, MSc (AEE INTEC)
DI Daniel Tschopp, MA BSc (AEE INTEC)
Lorenz Leppin, MSc (AEE INTEC)
Oliver Pimas (Know Center)
DI Dr. Max Toller, PhD (Know Center)
Christian Macher, MA CQF (Know Center)

The project *EnableDigitalDH* was supported by the Austrian Climate and Energy Fund and carried out as part of the Energy Research Programme (FFG 881 128). It addresses the submission focus areas „Data generation, provision and evaluation" (Datenerzeugung, -bereitstellung und -auswertung) and „Digitalization of integrated regional energy systems" (Digitalisierung integrierter regionaler Energiesysteme).

CONTENTS

Abstract

Background: District heating (DH) networks are considered an important technology for sustainable and economic heat supply in future energy systems, but their global renewable share is still around 6%. Present DH networks face the challenge to increase the renewable share and integrate with the electricity and transport sector. Digitalization in general and data driven analytics and machine learning (ML) methods in particular are seen as a very promising direction to meet these challenges and reduce the cost of decarbonization.

Problem definition: ML algorithms have been successfully applied for problems like load forecasting, fault detection, predictive maintenance and optimal scheduling. Their main focus is to learn intrinsic relationships from data and therefore, the created value crucially depends on the quantity and quality of the available data. Data requirements of data-intensive techniques have largely been neglected in the context of DH networks. The data quality of real measurement data is never perfect. The lack of available tools to handle anomalies and missing data is a bottle neck to make full use of ML techniques in the DH sector.

Aim of the project: The overall goal of the exploratory project EnableDigitalDH is to analyze and develop strategies to improve the robustness, precision and applicability of data-driven analysis, prediction and fault detection methods for DH networks. Hereby, the project focuses on an interdisciplinary approach.

Results: ML applications in the DH sector are well explored, and several review papers exist on the subject, but there is no study in relation to the quality and quantity of the data. A first-hand analysis of 63 papers on ML applications to DH networks showed, that about half of the publications in the field report data problems. Almost all of these papers explain the data processing in some detail. The mentioned methods for anomaly detection and gap filling are largely practical heuristics (e.g.; 60% of papers do simple linear interpolation for gap filling), rather than state-of-the-art data science methods. About half of the papers do not report data problems and do not explain the data processing in detail.

In the project, three methods have been developed which are available on software repositories: A data pre-processing pipeline (DPP) for anomaly detection and missing data imputation for univariate time series building on data science methods, and two physics-based approaches for missing data imputation for multivariate time series, namely heat transfer station simulation (HT_sim) and district heating network simulation (DHN_sim). DPP is a highly robust, stable and generally application tool which can be used as a standalone or in combination with HT_sim and DHN_sim. Compared to the methods prevailing in the district heating field, it marks a substantial improvement, as it makes a clear methodological distinction between continuous and binary data, incorporates season length and state-of-art anomaly detection and offers a variety of missing data imputation methods which are selected according to a Monte Carlo simulation. HT_sim and DHN_sim allow, compared to DPP, to fill longer missing data gaps and data channels which are permanently missing as they include domain knowledge from heat transfer stations

and district heating network respectively. A main advantage of the HT_sim approach is, that the process to set up the models for any DH network is straightforward as heat transfer stations are very similar across the globe. Also, HT_sim choses the appropriate model automatically depending on which data channels are missing. DHN_sim requires a full-scale DH network simulation for missing data imputation, which proved to have little additional benefit, given its complexity. Additionally, two innovative combined approaches were developed, the most promising is the combination of DPP and HT_sim (DPP_HT).

Recommendations and outlook: To make ML applications in the DH sector more traceable, a guideline and standardized reporting format of the deployed data processing steps would be highly desirable. Data processing should be made with Open Source tools whenever possible and the use of public data sets would improve collaboration on data-related issues and ease a further exploration the impact of poor data quality on ML methods. As time series data imputation as a scientific discipline is not as mature as one might expect, developments in the field should be incorporated to improve data preprocessing pipelines. Future research should work towards full automation of anomaly detection and missing data imputation by further developing the DPP approach and fine-tune it to specific boundary conditions. Furthermore, DPP should be combined with simulation-based bottom-up approaches for DH components which have a high degree of standardization in terms of typically available data channels and modeling properties, e.g., biomass boilers, oil boilers, gas boilers, solar thermal plants, heat pumps and water storages. For the HT_sim method, the development should focus on specific improvements like the handling of highly dynamic operating conditions, improvement of model parametrization and initialization and possible extensions to cases with more than two missing data channels. Full-scale DH network simulations for missing data imputation is not regarded as a promising direction for future research.

Kurzfassung (German)

Hintergrund: Fernwärmenetze gelten als wichtige Technologie für eine nachhaltige und wirtschaftliche Wärmeversorgung in zukünftigen Energiesystemen, aber der Anteil an eingesetzten erneuerbaren Energien liegt weltweit derzeit nur bei etwa 6 %. Fernwärmenetze stehen vor der Herausforderung, den Anteil der erneuerbaren Energien zu erhöhen und eine Koppelung dem Strom- und Verkehrssektor herzustellen. Die Digitalisierung im Allgemeinen und datengesteuerte Analysen und Methoden des maschinellen Lernens (ML) im Besonderen werden als vielversprechender Weg angesehen, um diese Herausforderungen zu bewältigen und die Kosten der Dekarbonisierung zu senken.

Problemstellung: ML-Algorithmen wurden bereits erfolgreich für Probleme wie Lastprognosen, Fehlererkennung, Predictive Maintenance und optimale Einsatzplanung eingesetzt. Ihr Hauptaugenmerk liegt darauf, intrinsische Beziehungen aus Daten zu lernen, und daher hängt der geschaffene Wert entscheidend von der Menge und Qualität der verfügbaren Daten ab. Die Datenanforderungen datenintensiver Techniken wurden im Zusammenhang mit Fernwärmenetzen weitgehend vernachlässigt. Die Datenqualität von realen Messdaten ist nie perfekt. Der Mangel an verfügbaren Werkzeugen zur Behandlung von Anomalien und fehlenden Daten ist eine große Hürde für die volle Potentialentfaltung von ML-Techniken im Fernwämesektor.

Ziel des Projekts: Das übergeordnete Ziel des Sondierungsprojekts EnableDigitalDH ist die Analyse und Entwicklung von Strategien zur Verbesserung der Robustheit, Präzision und Anwendbarkeit von datengesteuerten Analyse-, Vorhersage- und Fehlererkennungsmethoden für Fernwärmenetze. Dabei setzt das Projekt auf einen interdisziplinären Ansatz.

Ergebnisse: ML-Anwendungen im Fernwärmesektor sind gut erforscht, es gibt mehrere Übersichtsarbeiten zu diesem Thema, aber es gibt keine Studien in Bezug auf die Qualität und Quantität der Daten. Eine detaillierte Analyse von 63 Publikationen zu ML-Anwendungen für Fernwärmenetze ergab, dass etwa die Hälfte der Veröffentlichungen über Datenprobleme berichten. In fast allen dieser Veröffentlichungen wird die Datenverarbeitung erläutert. Bei den erwähnten Methoden zur Erkennung von Anomalien und Füllen von Lücken handelt es sich größtenteils um praktische Heuristiken (z. B. verwenden 60 % der Beiträge eine einfache lineare Interpolation zum Füllen von Lücken) und nicht um State-of-the-Art Data Science Methoden. Etwa die Hälfte der Beiträge berichtet nicht über Datenprobleme und erläutert die Datenverarbeitung nicht im Detail.

Im Rahmen des Projekts wurden drei Methoden entwickelt, die auf Software Repositories verfügbar sind: Eine Data Pre-Processing Pipelilne (DPP) für die Erkennung von Anomalien und die Imputation fehlender Daten für univariate Zeitreihen, die auf Data-Science-Methoden aufbaut, und zwei physik-basierte Ansätze für die Imputation fehlender Daten für multivariate Zeitreihen, nämlich die Simulation von Wärmeübergabestationen (HT_sim) und die Simulation von Fernwärmenetzen (DHN_sim). DPP ist ein sehr robustes, stabiles und allgemein anwendbares Verfahren, das als eigenständiges Programm oder in

Kombination mit HT_sim und DHN_sim verwendet werden kann. Im Vergleich zu den im Fernwärmebereich vorherrschenden Methoden stellt dies eine wesentliche Verbesserung dar, da eine klare methodische Unterscheidung zwischen kontinuierlichen und binären Zeitreihen vorgenommen wird, die Season Length und eine State-of-the-Art Anomalie-Erkennung integriert ist und eine Vielzahl von Methoden zur Imputation fehlender Daten vorhanden sind, die auf der Grundlage einer Monte-Carlo-Simulation ausgewählt werden. HT_sim und DHN_sim ermöglichen es im Vergleich zu DPP, längere fehlende Datenlücken und dauerhaft fehlende Datenkanäle zu füllen, da sie Domänenwissen von Wärmeübergabestationen bzw. Fernwärmenetzen enthalten. Ein Hauptvorteil des HT_sim-Ansatzes ist, dass der Prozess zur Erstellung der Modelle für jedes Fernwärmenetz einfach ist, da die Wärmeübergabestationen weltweit sehr ähnlich sind. Außerdem wählt HT_sim automatisch das passende Modell aus, je nachdem welche Datenkanäle fehlen. DHN_sim erfordert für die Imputation fehlender Daten eine vollständige Simulation des Wärmenetzes, was sich angesichts der Komplexität und des geringen zusätzlichen Nutzens als wenig praktikabel herausgestellt hat. Darüber hinaus wurden zwei innovative kombinierte Ansätze entwickelt, von denen der vielversprechendste die Kombination aus DPP und HT_sim (DPP_HT) ist.

Empfehlungen und Ausblick: Um ML-Anwendungen im Fernwärmesektor transparenter zu machen, wird empfohlen, einen Leitfaden und ein standardisiertes Berichtsformat für die angewandten Datenverarbeitungsschritte zu schaffen. Die Datenverarbeitung sollte nach Möglichkeit mit Open-Source-Tools erfolgen. Die Verwendung öffentlicher Datensätze würde die Zusammenarbeit bei datenbezogenen Fragen verbessern und eine weitere Erforschung der Auswirkungen schlechter Datenqualität auf ML-Methoden erleichtern. Da die Imputation von Zeitreihendaten als wissenschaftliche Disziplin noch nicht so ausgereift ist, wie man erwarten könnte, sollten die Entwicklungen auf diesem Gebiet zur Verbesserung der DPP einbezogen werden. Künftige Forschungsarbeiten sollten sich auf eine vollständige Automatisierung der Erkennung von Anomalien und der Imputation fehlender Daten hinarbeiten, indem der DPP-Ansatz weiterentwickelt und auf spezifische Randbedingungen abgestimmt wird. Darüber hinaus sollte DPP mit simulationsbasierten Bottom-up-Ansätzen von Fernwärmenetz-Komponenten kombiniert werden, die einen hohen Grad an Standardisierung in Bezug auf typischerweise verfügbare Datenkanäle und Modellierungseigenschaften aufweisen, z. B. Biomassekessel, Ölkessel, Gaskessel, solarthermische Anlagen, Wärmepumpen und Speicher. Für die HT_sim-Methode sollte sich die Entwicklung auf spezifische Verbesserungen wie die Handhabung hochdynamischer Betriebsbedingungen, die Verbesserung der Modellparametrierung und -initialisierung und mögliche Erweiterungen auf Fälle mit mehr als zwei fehlenden Datenkanälen konzentrieren. Vollständige Fernwärmenetz Simulationen zur Imputation fehlender Daten werden nicht als vielversprechende Richtung für zukünftige Forschung angesehen.

1. Introduction

1.1. Problem definition

Approximately 50% of the total final energy consumption worldwide is attributed to thermal energy including space and water heating, space cooling and industrial process heat [1]. As of 2018, only 10.2% of the energy supply of the heating sector was covered by renewable energy [1], which stresses the urgent need for the decarbonization of this sector.

An important technology for a sustainable and economic heat supply in future energy systems are district heating (DH) networks [2]. DH networks supply thermal energy to consumers through a network of pipes, connecting buildings and manufacturing sites in a neighborhood, village, town or whole city. The heat is generated from a centralized plant or distributed heat producing units [3]. Their main advantage is, that they can be incorporated into existing heating facilities with reasonable cost, using any available heat source such as combined heat and power (CHP) or waste-to-energy [4].

The global renewable share in district heating networks is still low (5.6 % in 2018) [1]. Present district heating networks need to transform to low-temperature networks with a high renewable share and integrate with the electricity and transport sector. These grids are oftentimes called 4th Generation District Heating Technologies and Systems (4GDH). They typically combine fluctuating renewable energy sources such as solar and wind with residual sources including waste and biomass to achieve high renewable shares [3]. Current research even points in the direction of 5th Generation District Heating and Cooling (5GDHC) and 6th Generation District Energy Systems (6GDES) [5]. The smart integration of different energy grids, storages, more and more decentralized renewable sources with intermittent production and varying temperature levels as well as higher demand on thermal comfort pose new challenges to the optimal operation of DH networks [6].

Digitalization of district heating networks helps to address these challenges and offers new opportunities to reduce the cost of decarbonization by optimization of operation, planning and business models [7]. In the broad sense, digitalization means the innovative use of information and communications technologies (ICT), in particular the large-scale rollout of smart devices and sensors, and the use of big data collection and analysis [7]. For district heating networks in particular, the use of data driven analytics and machine learning methods is seen as a very promising direction. Machine learning algorithms have been applied for problems like load forecasting, weather forecasting, fault detection, predictive maintenance and control and optimal scheduling [8]. Typical control challenges for which machine learning can improve existing solutions are unit commitment (the sequence of production units to start or stop), setting of optimal supply temperature levels and pressure heads on the supply side, demand side management on the substation level and optimization of the consumption behind the substation [9]. Their focus is on computational techniques to learn intrinsic relationships from structured and unstructured data, resulting in models that generalize these relationships for the analysis of previously unseen data [9].

The value created by data-intensive techniques crucially depends on the quantity and quality of the available data. There are regulations in place to support the trend towards higher data availability like the EU Energy Efficiency Directive [10], which mandates the supply of smart energy meter to final customers to accurately measure their energy consumption. However, up to the present the roll-out of smart meters has largely been limited to the electricity sector [7], and large scale roll-out of smart metering in DH networks has been limited to Scandinavian countries, China and a few others [11]. Additionally, it is essential to provide a suitable communications infrastructure for smart meter data storage and transmission. Low-cost data transfer infrastructures such as Long Range (LoRa), Narrowband IoT (NB IoT) are not fully rolled out at EU level [12]. Since smart meter data is personal data, it induces discussion of the need for individual consent from all customers. Accurate forecasts of the heat demand are impossible if half of the occupants in every building choose to deactivate their smart meter [12].

Overall, data requirements to apply data-intensive techniques have largely been neglected in the context of DH networks. In Europe, there exists no common standard regarding measurement equipment, data transfer and data storage for DH networks. In practice, the quantity and quality of the available data seems to depend largely on the age and degree of modernization of the network, production and consumption side, data policy of the energy supplier and grid operator, and country- and region-specific boundary conditions. Data issues are seldomly discussed in energy research and the energy sector as a whole is lagging behind regarding open data, such that it is even difficult to even assess the data availability [13].

The data quality of real measurement data in never perfect. The risk of losing data or anomalies in data, referred as bad data, are inevitable due to the sensor aging, communication delays, and physical damages of smart meters or components. It has already been shown for electricity meters that even advanced metering infrastructure systems fail to record around 2.7% of meter readings in a given year [12]. Problems like duplicates, error codes, strings, statistical outliers, physically implausible values, data gaps and other anomalies pose problems to machine learning algorithms, where they might be not applicable at all or produce misleading results [14]. Working with measurement data therefore requires data preprocessing and strategies to handle missing data for the application for data-intensive techniques for DH networks.

1.2. Aim of the project

The overall goal of the exploratory project *EnableDigitalDH* is to analyze and develop strategies to improve the robustness, precision and applicability of data-driven analysis, prediction and fault detection methods for DH networks. Hereby, the project focuses on an interdisciplinary approach to improve the data quality for data collected by DH networks. Specifically, the project revolves around the following key points:

- Systematic evaluation and characterization of data-driven models and their data quality requirements and associated limitations
- Evaluation of the possibilities, applicability and potential benefits of data science methods for measurement data of DH networks

- Exploration of the possibilities of a simulation-based data imputation approach to improve filling of longer data gaps or several simultaneously missing data channels
- Development of coupled methods for characterization and improvement of data quality with intelligent data-driven models
- Proof of concept of selected methods by means of executable implementation in a rapid prototyping environment
- Definition of the need for further research and creation of an Open Access Repository to initiate and support further research activities.

The project is conducted as part of the Energy Research Programme – 6th Submission. It adresses the submission focus areas „Data generation, provision and evaluation" (Datenerzeugung, -bereitstellung und -auswertung) and „Digitalization of integrated regional energy systems" (Digitalisierung integrierter regionaler Energiesysteme).

1.3. Methods

The project combines four main scientific methods:

- Literature and software review to systematically analyze the state-of-the-art of machine learning and data science methods used for DH data, including an analysis of strengths, weaknesses, opportunities and threats of different approaches and incorporating insights stemming from interviews of DH domain experts regarding practical difficulties and boundary conditions (see Section 2)
- Software implementation of a data imputation pipeline based on data science techniques for anomaly detection (using continuous/binary classifier, season length detection, multiple linear regression with time-based features, fourier series and lagged features, residual analysis) and different imputation methods (interpolation, moving average and season decomposition models) (see Section 3)
- Software implementation of a data imputation procedure using physics-based simulation models for a whole network and heat transfer stations (see Section 4)
- Design of two alternative combined data imputation methods using data science techniques and physics-based simulation models (see Section 5)

1.4. Structure of this report

The structure of this report is as follows. Section 2.1 gives an overview of the data collection in DH networks and the state-of-the-art for machine learning and data science methods applied to DH data. Section 2.2 describes the missing data imputation for univariate time series using data science techniques, whereas Section 4 describes missing data imputation for multivariate time series using physics-based simulation models. The combination the approaches of Section 2.2 and Section 4 are shown in Section 5. Section 6 contains the discussion, conclusion and outlook.

2. Data-driven methods applied to DH networks

In this Section, an overview is given on the data sources used in data-driven methods applied to DH networks, the state-of-the-art of machine learning methods and data science methods with regard to DH applications.

2.1. Data collection in DH networks

2.1.1. Data availability

The respective equipment or system for data acquisition, transmission and storage deployed in DH networks depends on the age and degree of modernization of the network and various other parameters like the number of consumers, internal standards of energy supply companies and country- and region-specific boundary conditions. The vast majority of central heat generation plants are equipped with Supervisory Control and Data Acquisition (SCADA) systems. The type of system or implementation also varies greatly based on the software, acquisition and transmission standards, measurement and storage intervals, recording periods, data export interfaces/formats. In Austria, high temporal resolution data (1 to 15 min) of all relevant operational parameters within the system are typically available for newer and larger systems. Smaller and older systems (e.g., biomass local heating systems built 20 years ago) may have limited data availability. The DH transfer stations at the heating customers can have very different measurement and control equipment and data acquisition with the lowest common denominator being the heat meter, which is also used for billing. For billing purposes, however, an annual reading is sufficient, which can be done on site, via radio or data cable. Only in the case of or special heat consumers or area stations, there are automated readings at shorter intervals (e.g., monthly) or even continuous measurement with automated data transmission. However, the trend for customer stations is also clearly moving towards continuous and highly time-resolved operating data acquisition. For example, in Vienna, all area stations and many large customers have already been equipped with detailed and continuous data acquisition and recording (standard for new large customers). The same applies to many other plants in Austria and also internationally. There are also already first major roll-outs of district heating smart meters in district heating networks in Scandinavia and China as well as Germany [15].

2.1.2. Data channels

Data-driven approaches use a variety of data sources relating to DH networks. As elaborated before, the quantity and quality of the available data varies greatly among different DH networks. Modern networks can have several thousand data channels with high resolution data. Below an overview is given of the different domains where data is collected.

A) Data for DH primary side and the secondary side at heat exchanger of heat transfer stations

A typical DH network can be divided in generation, transmission, distribution and thermal storage (Figure 1). The generation side consist of one or several heat producing facilities, which are connected to a transmission network and additional smaller distribution networks, operated with water as the medium of heat transfer, where the heat is finally transmitted to end-user. Thermal storages serve as heat sinks when demand is low and as heat sources for peak demand.

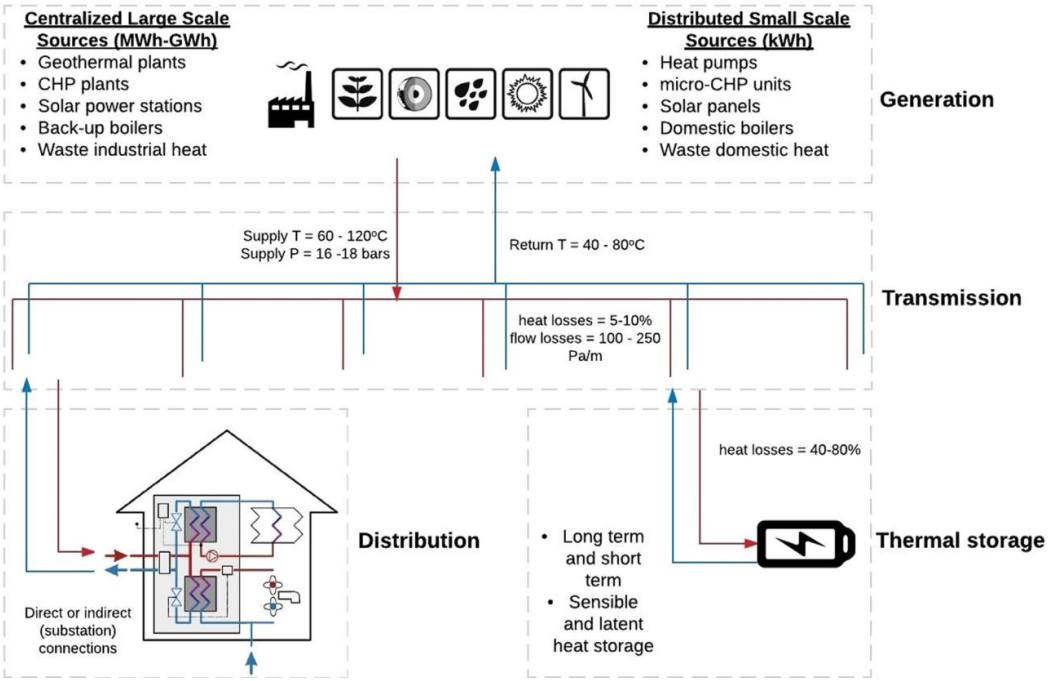


Figure 1: Generation, transmission and distribution of typical DH [6]

A first domain where data is collected is the primary side of DH networks (=generation side) and at the secondary side (=consumer side) at the heat exchanger of heat transfer stations. These data sources are distinguished from detailed data of the of generation and distribution side.

For the generation, typical data channels are

- Mass flow of heat source (e.g., gas boiler, biomass boiler)
- Supply temperature of heat source
- Return temperature of heat source
- Thermal power of heat source

For the tr
ansmission, typical data channels are

- Supply temperature of network
- Return temperature of network
- Pressure difference at feed-in point of network

For the distribution, heat transfer to end-user can be achieved by direct and indirect connections. In indirect connections, the primary and the secondary sides are separated by a heat exchanger. These connections are the most commonly used, as the primary side has the flexibility to operate at any pressure or temperature level without considering the local network of the heat consumer [6]. Heat transfer stations (substations) connect multiple consumers in parallel as shown in Figure 2. Typical data channels are:

- Primary side supply temperature of substation
- Primary side return temperature of substation
- Primary side mass flow of substation
- Secondary side supply temperature of substation
- Secondary side return temperature of substation
- Secondary side mass flow of substation

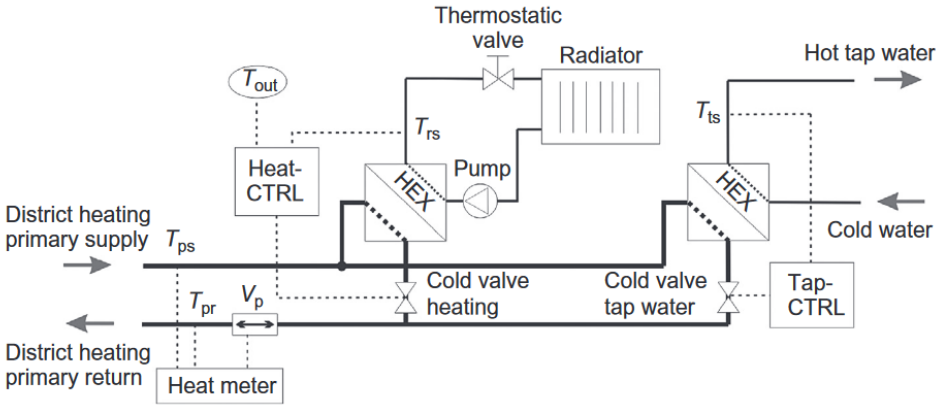


Figure 2: A Parallel coupled DH substation [16]

For thermal storages, typical data channels are

- Temperatures along storage height

B) Detailed data of generation side

The second domain where data is collected are heat sources on the generation side like gas boiler, oil boiler, biomass boiler, heat pump, solar thermal plant, etc. Each heat source has its own set of measurement data. As an example, typical data channels for gas boilers are supply temperature, temperature before return flow boost, temperature after return flow boost, temperature exhaust gas, oxygen percentage exhaust gas, demand signal, status signal, standby signal and alarm signal.

C) Detailed data of consumption side

The third domain are detailed data of the consumption side. These can include additional data directly measured near the heat transfer station like valve positions, pump signals, status signals, alarm signals, storage temperatures and data on heat circuits within a building like room temperature and occupancy levels.

D) Environmental data

The fourth domain are detailed environmental data like outdoor temperature, global and beam irradiance, wind speed, wind direction or relative humidity. These data are typically collected at the site of generation.

E) Socio-economic data

The fifth domain are socio-economic data which influence the heat demand of DH networks like non-working hours, public holidays or day of the week / weekend.

Data-driven approaches for DH use data from all of these domains. Therefore, data imputation techniques need to be very generic and focus on continuous data time series (thermal power, temperature measurements, etc.) as well as binary data time series (Alarm signal, status signals, etc.).

2.2. Machine learning methods for DH data

2.2.1. Overview of machine learning methods

This section provides an overview of ML methods which are commonly used for DH applications (see Section 2.2). The overview is based on [17], [18], [19], [20] [21], [22], [23]. Machine learning uses a variety of algorithms that iteratively learn from data. Figure 3 provides a graphical representation of the ML algorithms described in this overview. Statistical models in the narrow sense are described in Section 2.3.

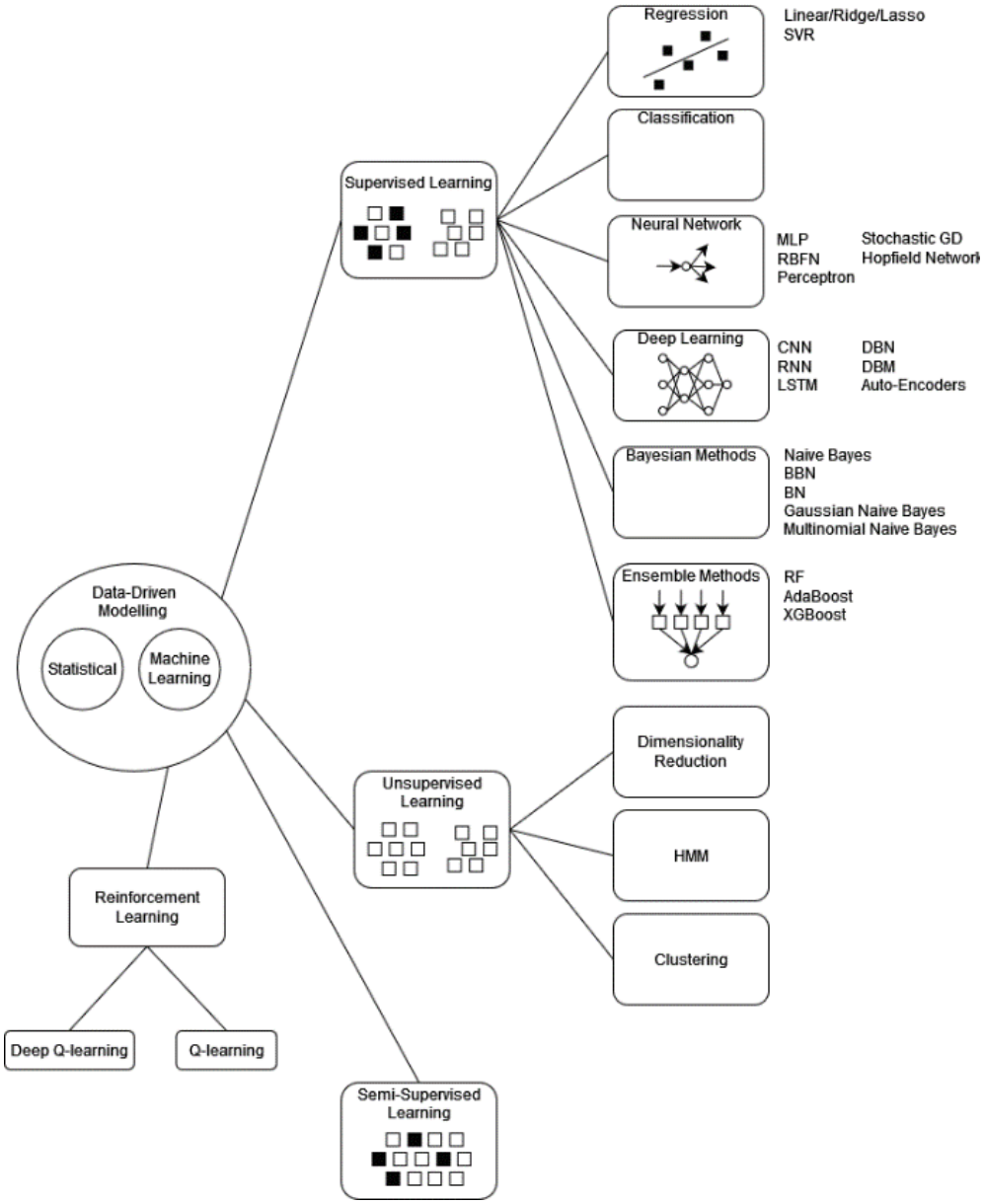


Figure 3: Overview of ML algorithms

In **supervised learning**, the learning algorithm is presented with a data set where the target variable is already known and the goal is to learn a rule that establishes relations between the inputs to their associated outputs.

- **Regression** methods iteratively refine the relationship between input and output variables using a measure of error in the predictions made by the model. Regularization algorithms such as ridge regression, fall under regression methods. These algorithms penalize models based on their complexity, favouring simpler models that are also better at generalizing. Support vector regression (SVR) is a regression application of support vector machines, which maximizes the margin between different categories. The goal is to find a linear regression function that could predict the result with acceptable deviation from the actual target. For nonlinear regression problems, a kernel

function should first be selected to map the original inputs to a high-dimensional feature space, and then apply the SVR. Therefore, one challenge of SVR is the proper selection of kernel function.

- **Support vector machines (SVM)** are binary classifiers that classify data instances by constructing a linear separating hyperplane. Even though they were initially used as classifiers, SVMs can be easily fit for forecasting, classification and clustering. With SVMs, it is possible to transform the original feature space in a higher dimensional representative hyperplane using nonlinear mappings (kernels), thus leading to enhanced classification capabilities.
- **Artificial Neural Network (ANN)** is a machine learning technique inspired by biological neural networks. A typical ANN usually consists of three layers: input layer, hidden layer and output layer. The training goal of an ANN model is to learn the weights and bias with proper number of neurons and hidden layers as well as activation functions. Note that although ANN with a single hidden layer can present any Boolean function and ANN with two hidden layers shows the ability to train any function to arbitrary accuracy, the number of hidden layers should be carefully selected to achieve better accuracy with fewer neurons. Wavelet neural network (WNN) is an algorithm based on the back propagation (BP) neural network topology, and the wavelet basis function is used as the transfer function of the hidden layer node. Extreme learning machine (ELM) for the feed-forward network of a single hidden layer. This method has several advantages. Firstly, it randomly generates the weights and thresholds of the input layer in the process of training and the determines the weights of the hidden layer and the output layer. Therefore, iterative calculations are not necessary and the unique optimization equation can be obtained by entering the number of nodes of the hidden layer. Secondly, ELM has a high learning speed and good generalization ability.
- Once the number of hidden layers is increased, the ANN could be considered as **deep learning**. Models that fall under deep learning are deep neural network (DNN), convolutional neural networks (CNN) and recurrent neural networks (RNN). A DNN is a complex version of ANN containing multiple hidden layers between input and output layers. Typical DNN is a feed-forward network without lopping back. Generally, DNN refers to fully connected networks, which means that each neuron in one layer receives information from all neurons from previous layer. The motivation to use DNN are that it requires less neurons than simple ANN in representing complex tasks, and that in practice DNN generally have higher prediction accuracy than ANN. However, DNN models suffer from overfitting and they are computing intensive. Compared with general DNN, CNN decreases the risk of overfitting by reducing the connectedness scale and structure complexity. Therefore, CNN could also be treated as a regularized version of typical DNN. CNN is well-known in the field of visual imagery analysis, such as image recognition, image classification, medical image analysis and natural language processing. The distinction between RNNs and other deep learning algorithms is that RNNs involve loops in their structure and make it possible that information flows in any direction. These cycles introduce time delay in RNN and make RNN more suitable to exhibit temporal dynamic behaviour. However, as the weight for the loop is the same for each time step, gradients in the traditional RNN tend to explode or vanish when the loop runs for many times. This problem is called long dependency. To solve this problem, one commonly

utilized RNN model, called Long Short-Term Memory (LSTM), could be applied to remember information for a long period.

- **Bayesian methods** construct models based on Bayes' Theorem. Common methods include Naive Bayes, Gaussian Naive Bayes and Multlinomial Naive Bayes and Bayesian Belief Networks (BBN).
- **Ensemble methods** are a category of algorithms that combine several machine learning methods, in order to enhance overall performance of the prediction. Bagging/bootstrap aggregating predicts the output by training the same baseline models parallel on different sub-datasets, which are sampled from original input datasets uniformly by replacement. This algorithm is used to reduce variance/overfitting when running the trained model on the validation set. Random forests (RF) are the most common bagging method. Boosting trains the baseline models iteratively. In each step the successive model tries to fix the mistake made by previous models by increasing the weight to observation which have been predicted incorrectly. AdaBoost and XGBoost (Extreme Gradient Boosting) are the common in boosting. In stacking, the idea is learning the problem with different types of models which are capable to learn some part of the problem, but not the whole space of the problem. It is applied on an arbitrary set of models. Different models are trained on the available input dataset, and then a meta-model is trained based on the outputs of these models to make the final prediction.

In **unsupervised learning** algorithms do not use output data that are different from the input, there is no separation between training and test data sets, and data is unlabelled. The main goal is to unveil features and hidden patterns in data or compress data into a more compact forms without significant loss of knowledge. Clustering is the task of organizing data in such a way that similar objects are placed into related or homogeneous groups without prior knowledge of the groups' definitions. Time-series clustering is a special case of cluster analysis that has been used in many scientific areas to discover interesting patterns in time-series datasets such as smart meter datasets. There are generally three different approaches to cluster time-series: the feature-based, model-based, and shape-based approaches.

In **semi-supervised learning** input data is a mixture of labelled and unlabelled examples. There is a desired prediction problem but the model must learn the structures to organize the data as well as to make predictions. Mainly the problems are classification and regression. Example algorithms are extensions to other flexible methods that make assumptions about how to model the unlabelled data.

In **reinforcement learning**, the learning algorithm dynamically re-adjust the output in order to maximize the notion of cumulative reward. Examples for reinforcement learning are Deep Q-learning and Q-learning.

2.2.2. Machine learning methods used in DH

The precision of the ML models mainly depends on the quality and quantity of the data used. In the field of district heating and cooling systems, there are many ML approaches in use for different purposes. With the aim of understanding the purpose and the type of machine learning methods applied on DHS in relation

to the quality and the quantity of the measurement data from these systems, a systematic literature review has been conducted.

The starting point of the literature review was an analysis of existing review papers in the field. These papers were selected using a constructed string search in the Scopus database *("Review" AND ("Data-Driven" OR "Machine Learning") AND "District heating")*. This search provided three review papers, published between 2018-2022, namely by Mbiydzenyuy et al. [9], Ntakolia et al.[8] and Buffa et. al. [24]

Ntakolia et al. [8] reviewed 74 articles from the period 2002 – 2019 for ML applications on the district heating and cooling sector. They categorized these articles according to the main application and the used ML models as shown in Figure 4.

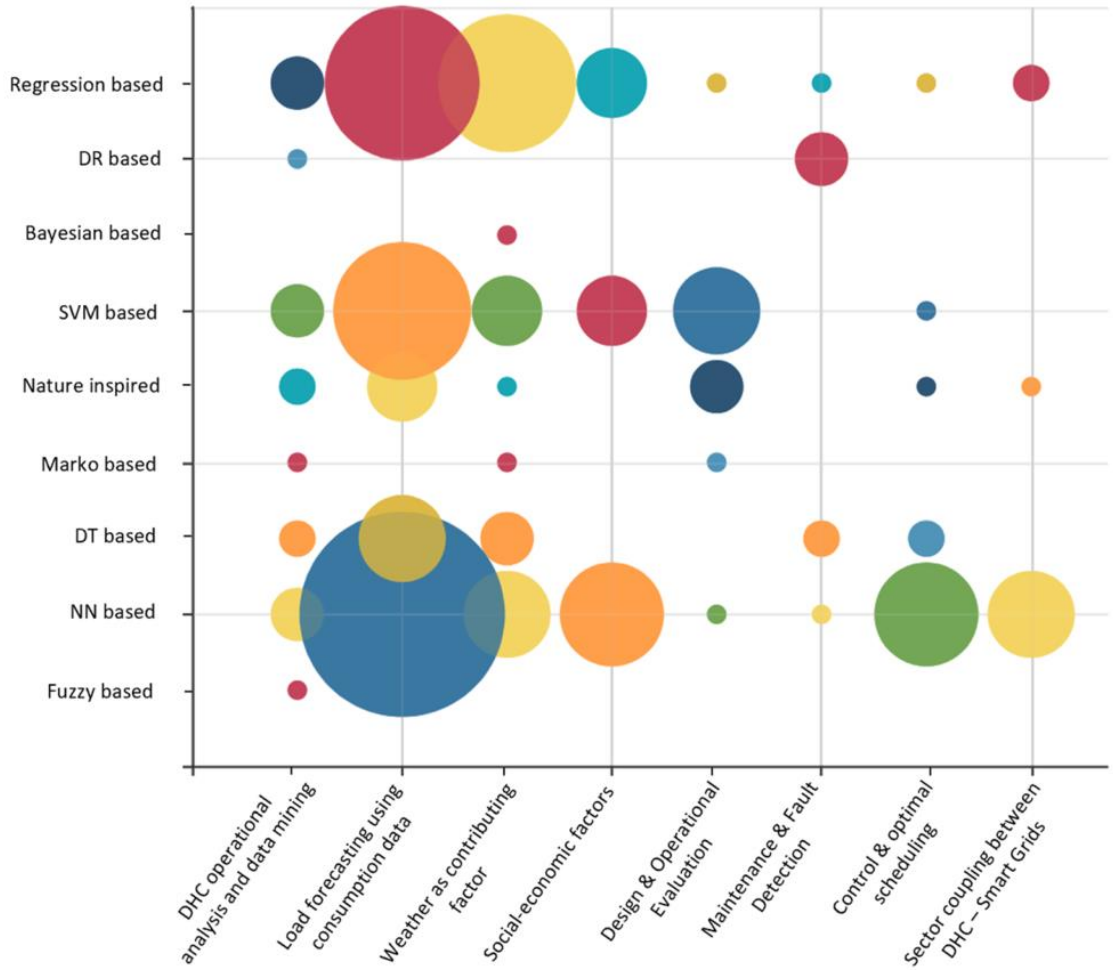


Figure 4: Bubble chart showing used ML techniques and main applications in the DHC sector [8]

ANN is the most, and regression based / SVM approaches are the second most, used model in both heat load and weather forecasting. Most of the DHC systems aimed at forecasting short-term load demands

(e.g., within 24 h) by correlating weather with historical load variables. Ensemble models, falling in multiple categories are the most used in fault detection in DHC substations. The main application of load forecasting using consumption data motivated by that fact that planning the energy consumption in advance and securing the supply to the demand is crucial for creating reliable systems as well as improving the energy efficiency. To plan the energy consumption, accurate prediction methods are required. ML models are efficient to predict the energy demand based on the building historical data of energy consumption without requiring any physical information (building envelop material, wall thickness, etc.) of the buildings.

Mbiydzenyuy et al. [9] reviewed 179 papers for ML applications on district heating and cooling . 72% of the reviewed papers fall under load forecasting. The second most popular category is anomaly detection. Anomaly detection is essential in the DHS systems to provide reliable and efficient energy supply. A fault in the DHS can be a pressure drop from the pumps, equipment deterioration or leakage in the system. Leakage faults in DHS can imperil the reliability of the system, decrease heating efficiency and increase the cost losses. In this context, maintenance of the DHS components is important as well, as the reliability of sensors is necessary for the correct billing purposes or operational decisions. The paper doesn't classify each ML method for each application but only discusses to what extent state-of-the-art ML has addressed key challenges in the DH industry, and identify discrepancies between the two fields to propose a road-map with suitable goals on how to increase the impact of ML in DH.

Buffa et al. [24] focuses on innovative methods on advanced control and fault detection strategies for district heating and cooling systems. Innovative methods are divided into machine learning and physical models. ML applications in the field of control and fault detection consist of heat load forecasting, optimization problems and anomaly (fault) detection. This study provides few examples on ML applications on heat load forecasting, optimization problems and anomaly detection but does not provide a systematic and comprehensive overview of these ML applications. The highlight of the paper is the increasing popularity of ML techniques on different applications due to its ability to cope with system non-linearities, and meta-heuristic optimization algorithms. According to this study, even though more algorithms based on ML techniques used for fault detection, the behaviour of the component forecasting of some of these approaches are not useful, since models trained with laboratory data or data coming from simulations do not achieve a good enough performance when working with online data or measurement data.

Based on these review papers, ML is mostly used in load forecasting, since the prediction of building energy usage plays a vital role in developing a model predictive controller for consumers and optimizing energy distribution plan for utilities. Therefore, a recent review study of Sun et al. [25] on building energy prediction has also been used to include relevant information from this field. This work reviews 105 articles on building energy prediction, covering the entire data-driven process, addressing data sources, feature types, model utilization and prediction outputs. 17% of the reviewed studies are public datasets and the rest are private datasets, which are not published due to privacy and ethics reasons. Meteorological information, historical data and time index are the most important factors for building energy prediction.

More than 60% of studies predict the energy for an entire building, followed by around 20% of studies for district level. The high number of studies on these two scales is caused by more collectable data and applicability of the developed model to realistic demand response control and grid distribution. On the other hand, the prediction for sub-building level is lacking due to the limitation of data collection. In term of prediction horizon, one paper could present results for several temporal granularities. Most studies (65%) present multi-step prediction. Besides, the resolution and horizon for most studies are higher than 1 min. ANN, SVR and LR seem to be the popular models, while the concentration on time series analysis and RT is less. The application of RT is less due to its unacceptable prediction errors when applied to validation or test datasets. Besides these methods, deep learning has started to draw interest in recent years.

2.2.3. Data problems and data sources

In the four review papers presented in the previous Section 2.2.2, detailed information on the ML application in relation to the quality and quantity of the dataset could not be found. Therefore, a systematic literature review with this focus has been conducted. The databases Scopus, Web of Science and IEEE Xplore were used. The focus of IEEE Xplore is electrical and computer engineering, which has been included since many ML algorithms are related with computer engineering domain.

The chosen strings for the search in the databases are shown in Section 2.2.2. In all databases, searches were done with the chosen strings set as “Keyword” (for Scopus additionally in “Article Title” and “Abstract”) and limited to articles published between 2017 and 2022. After duplicated papers were removed, 226 papers remained. Out of these, 63 papers (28%) were selected due to high relevance to the research question based a read of the abstract.

Table 1: Search strings for literature review on ML methods

#	**Strings**
1	“Machine Learning” AND “District heating”
2	(“Machine Learning” OR “Artificial intelligence” OR “Data-driven” OR “Deep Learning” OR “Neural Network”) AND (“District heating” OR “District cooling” OR “Thermal network”)
3	(“Machine Learning” OR “Artificial intelligence” OR “Data-driven” OR “Deep Learning” OR “Neural Network”) AND (“District heating” OR “District cooling” OR “Thermal network”) AND (“Energy” OR “Smart energy”)

Figure 5 illustrates the distribution of the type of data used in the reviewed papers and the distribution of the main applications. 76% of papers use measurement data, while 24% of the papers use simulation data. There was no control and design optimization application found for the measurement datasets and no clustering or pattern recognition application for datasets from simulation.

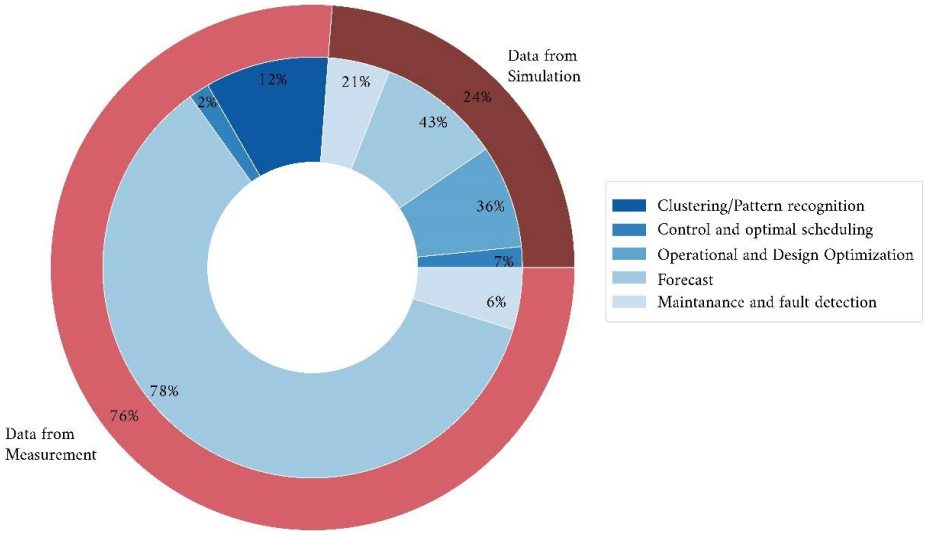


Figure 5: Distribution of data source (measurement vs. simulation) and share of application per data source for 63 reviewed papers.

The papers that use measurement data are classified separately in Figure 6. 79% of the papers belong to forecasting category, which coincides with findings from the presented four review papers. As can be seen, about half of the articles report data problems (dark orange colour in middle circle) and almost all of these papers explain the data processing in detail (dark green colour in inner circle). About half of the papers do not report data problems (light orange colour in middle circle) and do not explain the data processing in detail (light green colour in inner circle). It can be assumed, that data scaling, which is normalizing the input to improve training performance and reduce the problem of outliers, is performed almost in all papers because any average input, that is not close to zero, will result in slowing down learning. Most likely, a substantial share of the papers which do not mention data problems therefore does some data processing, but does not mention it, which makes the results hard to trace from a scientific point of view.

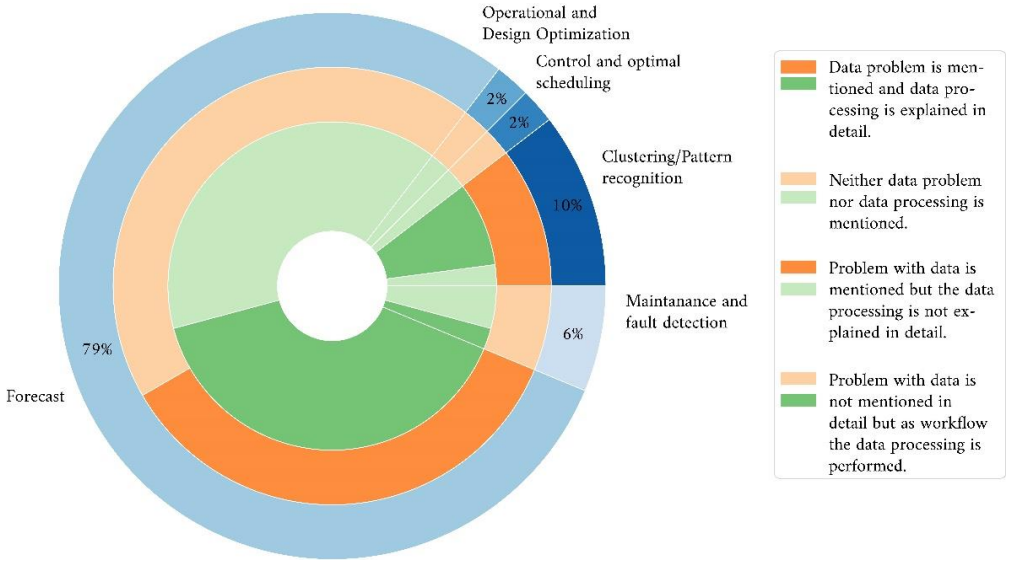


Figure 6: Reported data problems (orange circle) and explanation of data processing (green circle) for different applications (blue circle) for 63 reviewed papers.

Data problems which the reviewed papers mention are inconsistent data (e.g., the heating load value is zero when the heat pump is in operation), abnormal data (e.g., the heating load value is over the maximum load supply), missing data and duplicated data. The used methods to detect outlier/noise in the data and to fill data gaps are listed in Table 2: Table 2. When compared to the methods of Section 2.3 and Section 3, it seems that the deployed methods are mainly practical heuristics and do represent the state-of-the-art of data science techniques. Scientific methods for missing data imputation for smart electricity meter data (e.g.; [26], [27]) are not commonly used in the field.

Table 2: Methods to address data problems for 63 reviewed papers

Methods to detect outliers/noise	**Papers**
Local outlier factor (LOF) method	[28], [29]
Moving median method (MAD)	[30], [31], [32], [33]
Z-score outlier removal	[34]
Hampel filter	[35]
Three-sigma method	[36]
Methods to fill data gaps	**Papers**
Linear interpolation	[28], [30], [31], [33], [37], [38]
Cubic spline interpolation	[32]
Quadratic interpolation	[35]
Filling the shortest gaps with averaged values of respective measurements from the closest vicinity	[39], [36]
Other methods	
Kalman filter for data pre-processing	[40]

One of the few studies addressing the effect of missing data in the DH sector claims, that missing data of 1-7 days within a 3-months period has little impact on the prediction of heating energy consumption of DH station with MLR, RF, ANN and RNN models [41]. However, these results are hard to generalize as the effect of missing data on the performance of ML method depends on the various conditions like the pattern of missing data (Missing completely at random (MCAR), Missing at random (MAR), Not missing at random (NMAR) or Non-ignorable), the ML method, characteristics of the data set (time series, continuous or discrete data set), missing percentage or test data length [42]. A common way to handle missing data by simply ignoring or deleting them would reduce the amount of data and lead to a nonresponse bias, that can be moderated by using advanced data science techniques.

2.3. Data science methods for DH data

In the data science literature, there are several interesting and novel state-of-art methods that can be applied, adapted, or extended for the specific needs of data-driven analyses for DH networks. Since district heating data are frequently stored in a sequential, temporal format in which erroneous data are common (e.g., outliers, anomalies, missing data), the respective relevant sub-areas of data science literature are time series data mining and anomaly detection. The state-of-the-art of these two techniques and a further

novel state-of-art technique that is likely highly relevant for enabling digital district heating, namely parameter free clustering, is elaborated in this Section. A working paper on the current state of the art was prepared during the course of this project which is available at: https://doi.org/10.5281/zenodo.7470102.

2.3.1. Time series data mining methods

Time series data mining is a challenging task that is actively researched by numerous scholars. One main reason why time series data mining is more challenging than traditional data mining is that time series data are not necessarily statistically independent from their past (or future), and that the underlying distributions may also change over time. However, this inherent statistical dependence also has an advantage since it allows a construction of statistically meaningful forecasts of future data, e.g., future energy consumption, by establishing a comprehendible relation between successive data points.

Time series forecasting methods have multiple applications in data-driven DH which range from optimizing local efficiency [43] over construction of realistic energy policies [44] to merely filling up missing data points caused by unreliable sensors [45]. One of the most well-known models for time series forecasting is the ARMA/ARIMA model. The model is composed of two parts: an autoregressive part (AR) and a moving average part (MA). ARMA could only handle stationary time series. When predicting nonstationary time series, ARIMA would be a better choice since it integrated an initial differencing step to eliminate the non-stationary. ARMA and ARIMA show the ability to consider the effect of historical data, thus, their prediction performance would be acceptable if the output is highly impacted by previous values. However, determining the orders for AR and MA models and the times of initial difference would be a challenge.

Besides ARMA/ARIMA, several more traditional models that may also have potential value for district heating, namely Exponential Smoothing (ES) and the Autoregressive Model (AR). Moreover, the machine learning community has devised several time series data models based on neural networks that are amenable to forecasting, namely Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM), and Gated-Recurrent Units (GRU). Altogether, the statistical models ES, AR, and ARIMA belong to the same hierarchy as the machine learning models RNN, LSTM, and GRU. Choosing a higher-level model such as LSTM over a low-level model such as ES has the advantage that more complicated time series data can be predicted, at the cost of increased complexity of model training and hyperparameter selection. However, it must the stressed that each of the aforementioned models is likely sufficient for forecasting any time series data that were collected in the context of digital district heating as long as the underlying series is stationary. It is more critical to select appropriate preprocessing techniques which allow one to transform district heating time series data into a stationary form.

To succeed at such preprocessing several deterministic components need to be extracted from the raw data, e.g., trends, changepoints, and seasonality. However, it must be noted that the state-of-art in this sub-area of data science is less explored than forecasting. Hence, in practice it is mostly necessary to design application-specific solutions that are tailored for the time series data at hand.

2.3.2. Anomaly detection methods

Suspicious data that deviate considerably from other data, also known as anomalies, are particularly common in real-world sensor data. There are many typical causes for such anomalies, e.g., sensor malfunction, data transmission subject to heavy-tailed noise, and transmission errors. However, what most anomalies have in common is that they complicate streamlined data processing and hinder successful automation regardless of the cause of the anomaly. This explains why detecting anomalies is crucial in numerous really world applications. Currently, the state-of-art of anomaly detection is defined by four methods (for an overview see [46]):

- **Isolation Forest** [47]. Anomaly scores are inferred using random axis-parallel subdivision to isolate each data point in a separate hyper-cube. Key advantages of this method are that it is intuitive, can be easily visualized and has linear computational complexity. Drawbacks of this method are its lack of a theoretical foundation and that the difficulty of fine-tuning the hyperparameters.
- **Robust Kernel Density Estimation** [48]. Anomaly scores are indirectly proportional to a data point's density as per a kernel density estimate that was inferred using robust statistics. This method is builds upon well-understood statistical theory and is flexible. However, this method has quadratic computational complexity, which greatly affects scalability.
- **Subspace Outlier Detection** [49]. Anomaly scores are the standard deviation of a low-dimensional subspace that contains a data point and its nearest neighbors. This method has a distinct advantage over many others in high-dimensional datasets. However, its value is disputed for low-dimensional datasets.
- **Dirichlet Process Mixture Models** [50] Anomaly scores are inferred using variational inference of a mixture of Dirichlet distributions. This method builds upon well-understood statistical theory, yet its computational complexity is abysmal.

Given their methodological diversity, it seems promising to take all four methods into one's pre-processing method arsenal. It must be noted that these methods were primarily designed for datasets consisting of independent and identically distributed data. Since time series data, which are not independent and identically distributed, are commonly recorded in district heating, it may be necessary to resort to dedicated detectors specialized on time series data. Unfortunately, most of these dedicated methods have recently been shown to be unreliable [51]. Therefore, it seems advisable to currently stick to the four above-mentioned well-understood methods to treat data of DH networks.

2.3.3. Parameter-free clustering

Beyond time series data mining and anomaly detection methods, there is a further novel state-of-art technique that is likely highly relevant for DH data: Efficient Parameter-Free Clustering using First Neighbor Relations (FINCH) [52]. FINCH is a clustering technique that constructs an easily interpretable hierarchy of logarithmically many clustering of a dataset without the need of hyperparameter tuning. This is beneficial for many different steps since clustering techniques allow the discovery of classes and assign data points to these classes for almost arbitrary datasets. For example, if one seeks to categorize buildings according to their energy consumption, then two steps are necessary:

1. Discover meaningful classes, e.g., high consumption vs. low consumption

2. Determine for each building to which discovered class it fits best

FINCH can perform both steps in sub-quadratic time. However, the main advantage of FINCH is arguably not its efficiency but its parameter-freedom. The technique just requires an arbitrary dataset as input and constructs logarithmically many clustering of this dataset. Because of the logarithm number of clustering, it is easy to manually select which clustering is best suited for the underlying context. Overall, FINCH seems to be a promising technique that likely has additional uses in district heating that are waiting to be discovered.

3. Anomaly detection and missing data imputation using data science techniques

3.1. Overview and preparatory steps

Building on state-of-the-art data science methods (see Section 2.3), a pre-processing pipeline for univariate time series data (in short DPP data pre-processing pipeline) has been developed. The aim of the pre-processing pipeline is to get a highly robust, stable and generally applicable tool for anomaly detection and missing data imputation. The pipeline covers both continuous and binary time series data and therefore the two main data categories which data-driven methods in the context of DH networks rely on (see Section 2.1). Compared to commonly used anomaly detection and missing data imputation for machine learning applications the context of DH network (see Section 2.2), the developed pipeline is a major methodological improvement. The pre-processing pipeline is available as a software programme written in R with an extended documentation and a description of use cases (see Appendix A 8.1), this Section describes the main building blocs.

Figure 7: Data pre-processing pipelineFigure 7 shows a flow-chart of the pre-processing pipeline. To meet the requirement of a robust pipeline, features have first be classified into continuous and binary data. In order to classify the data according to the categories, a rule-based approach has been implemented as follows:

- Condition 1: Check if the length of distinct values of the feature (without missing values (NA)) is greater than 3
- Condition 2: Check if the two most common values have a share less than 99.5%
- If Condition 1 and 2 are TRUE than the feature is classified as continuous, otherwise as binary

Continuous data will be treated differently in anomaly detection and imputation than binary data. This is the reason why it has to be separated. After classifying the data type, every feature will be prepared independently.

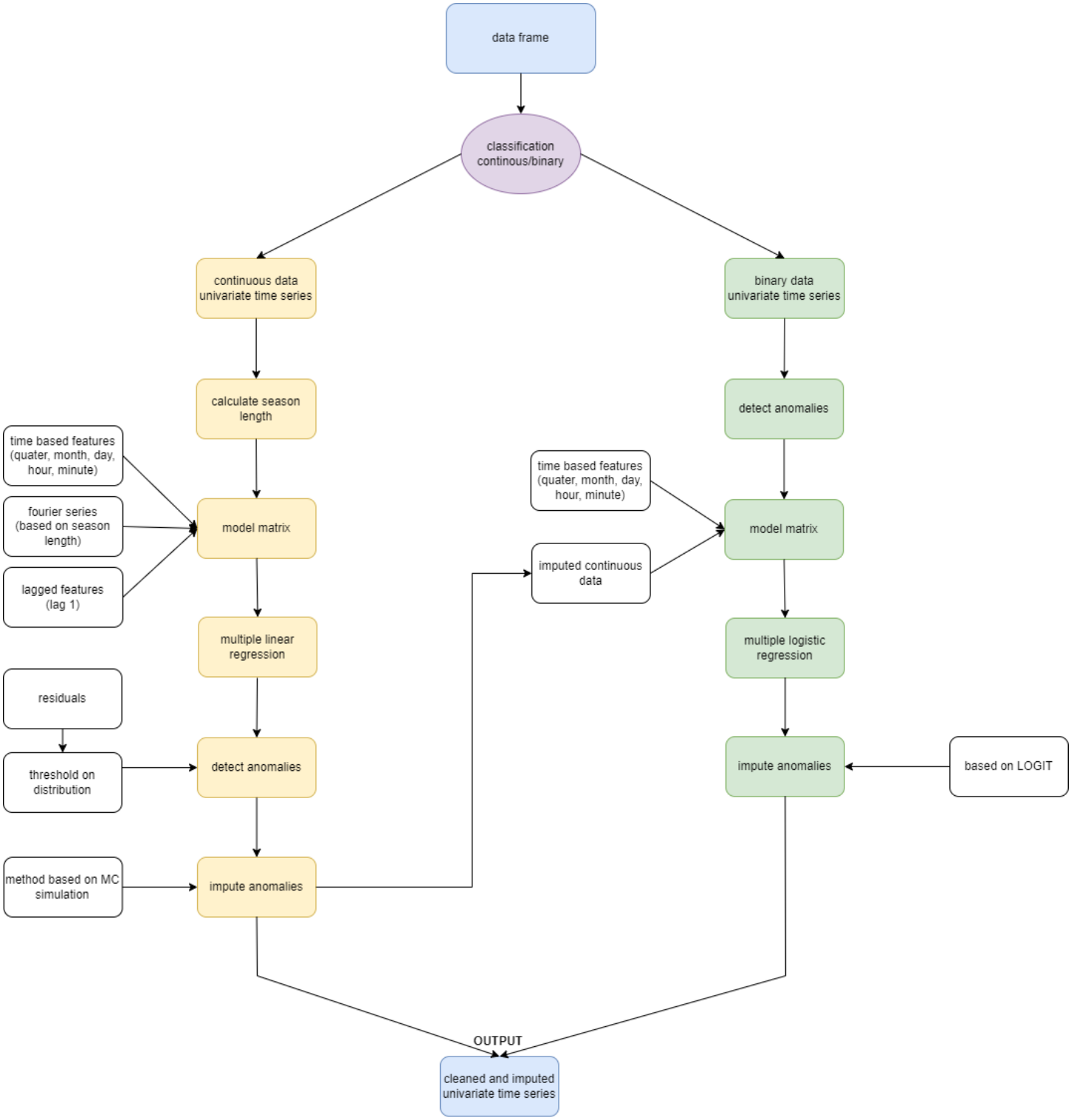


Figure 7: Data pre-processing pipeline

3.2. Pipeline for continuous data

Season length: One of the most important information of a continuous time series is the season length. In this process step an attempt is made to determine the season length automatically based on a

combination of three different estimates computed on the feature time series and its 10-fold-self-composed autocorrelation [40]. The default value for no determined season length is 1. Only time series with a season length greater than 1 will be subsequently used for anomaly detection and imputation.

Model Matrix: For continuous data, anomalies will be detected based on a multiple linear regression method. After calculating the season length, a model matrix will be created for applying the regression model. The model matrix consists of three groups of features which will be generated automatically.

- *Time* based *features:* Based on the time index separate information about quarter, month, day, hour and minute will be extracted. In addition, the time index will be converted to a numerical feature.
- *Fourier series:* Based on the previous calculated season length, Fourier terms will be calculated. Fourier terms will be used to cover seasonal patterns [41].
- *Lagged features:* Lags are very useful in time series analysis because of autocorrelation, which describes a tendency for the values within a time series to be correlated with previous copies of itself. For creating the model matrix only lag 1 will be used.

Multiple Linear Regression: For continuous time series data, a multiple linear regression will be used for anomaly detection. The reason for this is to get a generally applicable method if exploratory data analysis showed that the non-random components are likely "Gaussian". We assume that residuals are independent and identically distributed with $N(0, \sigma^2)$.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \qquad \text{Eq (1)}$$

$y_i = \text{dependent variable}$
$x_i = \text{explanatory variable}$
$\beta_0 = \text{y\_intercept}$
$\beta_p = \text{slope coefficients for each explanatory variable}$
$\epsilon_i = \text{error term}$
$\epsilon = iid\ N(0, \sigma^2)$

The dependent variable is one univariate continuous feature (y_i) and the independent variables are the features which has already been defined in the model matrix (x_i).

Detect anomalies: In this pipeline the anomaly detection uses a probabilistic approach based on residuals from the multiple linear regression. In the first step a lower and upper bound will be defined. Anomalies are points which are above upper bound or below lower bound as shown in Figure 8.

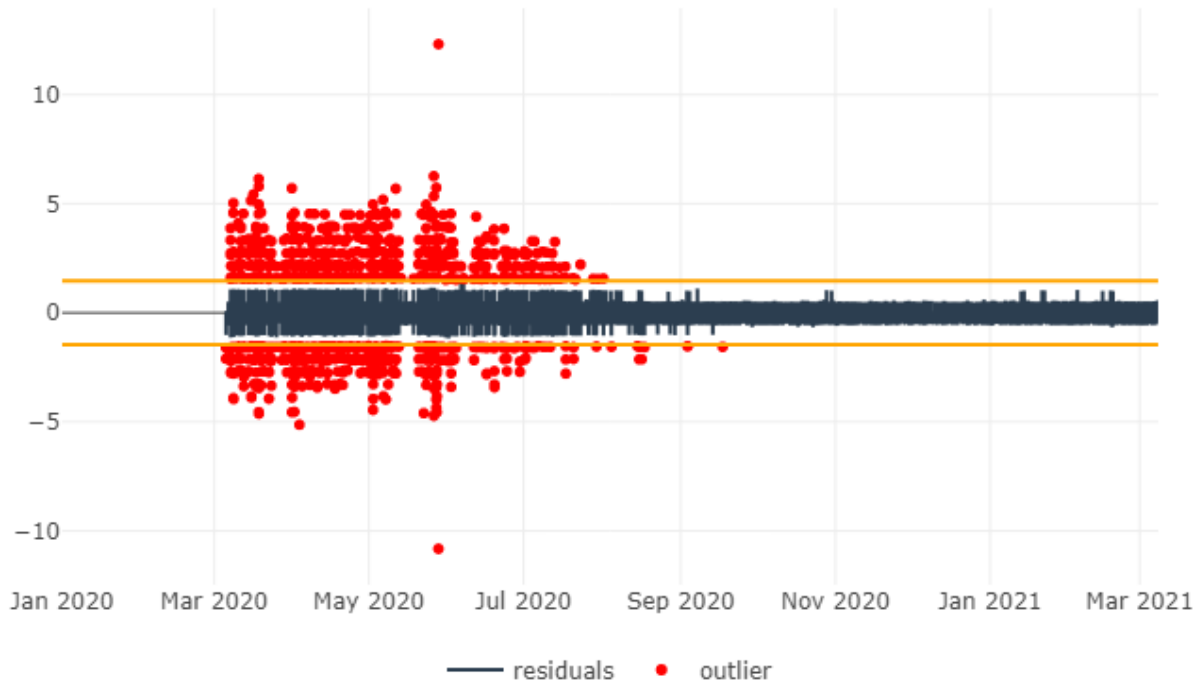


Figure 8: Residuals, lower and upper bound and outliers for anomaly detection

Impute anomalies: In the imputation process step all detected anomalies and missing values (NA) will be replaced by values from a specific imputation method.

Imputation Methods

- *Interpolation Methods*
  - Linear interpolation
  - Spline interpolation
  - Stine interpolation
- *Moving Average Methods: For* moving average algorithms, a rolling window of 10 and 20 were used.
  - Simple moving average (SMA)
  - Linear weighted moving average (LWMA)
  - Exponential weighted moving average (EWMA)
- *Season Decomposition Methods:* This method removes the seasonal component from the time series, performs the selected imputation method (interpolation) on the de-seasonalized series and afterwards adds the seasonal component again.
  - Seasonally decomposition linear interpolation
  - Seasonally decomposition spline interpolation
  - Seasonally decomposition stine interpolation

Selection of Imputation Method

The method which will finally be applied will be selected based on a Monte Carlo simulation as shown in Figure 9. In order to create the Monte Carlo simulation two parameters are necessary and must be set in advance.

- n_samples
  - The number of how many existing values/entries are replaced by NA
  - default: 100
- N_SIMULATION
  - The number of simulation loops
  - default: 5

The first step is to replace existing values (randomly selected) from the univariate time series (outliers have already been removed in the previous step) with NA. The number of replacements will be set with the parameter n_samples. After creating NA all types of imputation methods are applied and for every method the mean absolute error (MAE) will be calculated. This process would be one simulation (N_SIMULATION).

Finally, the lowest average MAE of all simulations illustrates the most suitable imputation method which will be used for replacing anomalies. At the end two results will be produced.

1. Partially imputed time series
   All anomalies and NA will be replaced except a predefined maximum gap size
2. Fully imputed time series
   All anomalies and NA will be replaced

Imputation Gap

If missing values will be imputed it make sense to limit the maximum gap size which has to be filled by the method, otherwise imputing huge missing gaps could lead to distorted results and inaccuracies. The maximum gab size will be determined based on existing data points per day.

1. Determine the number of data points per day
2. Filter those number of data points which have the majority of available days
3. Maximum gap size is 25% of that number of data points

Figure 10 provides an example to determine the maximum gap size for the imputation: (a) The data consists of a continuous time index. (b) Based on the date the number of data points (n_Day) will be calculated. (c) Finally, a counter (N) will determine the distribution of the retrieved number of data points. In the example, the value n_Day = 288 is counted 430 times and has the highest frequency. Therefore the maximum gap size is 72 (288 x 25%). Gaps with more missing values than 72 will not be imputed.

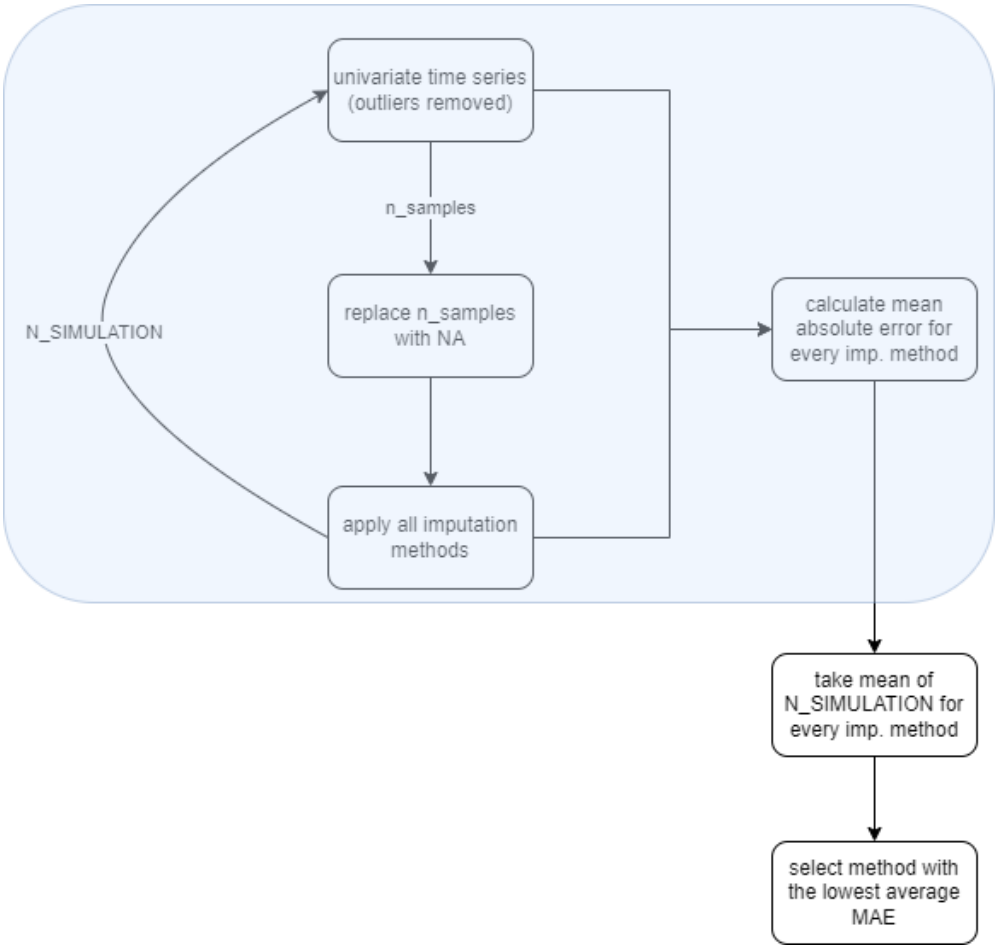


Figure 9: Monte Carlo Simulation - Imputation

Time_index
2020-01-01 00:00:00
2020-01-01 00:05:00
2020-01-01 00:10:00
2020-01-01 00:15:00
2020-01-01 00:20:00
…
….

(a) time index

→

Date	n_Day
2020-01-01	288
2020-01-02	288
2020-01-03	288
2020-01-04	288
…	…
…	…
…	…

(b) number of data points per day

→

n_Day	N
84	1
276	1
283	1
288	430
300	1
…	…
…	…

(c) distribution of n_Day

Figure 10: Example to determine the maximum gap size for the imputation

3.3. Pipeline for binary data

For all univariate continuous time series, the process from calculating the season length to anomaly imputation is run before starting the pipeline for binary data, because their input will be used in the model matrix.

Detect anomalies: For binary data a simple approach based on percentage occurrence is used for anomaly detection. For every binary feature the distinct values/entries are counted and their percentage is calculated. If some value is not one of the two most common entries in the table, this value will be highlighted as an anomaly, see Table 3 for an example.

Table 3: Anomaly detection for binary data. Value „99" is not among the two most common entries and therefore highlighted as an anomaly.

Values	N	Percentage	Valid
0	124 475	99.978%	Yes
1	25	0.020%	Yes
99	2	0.002%	No

Model Matrix: For binary data, anomalies will be detected based on a multiple logistic regression method. After preparing continuous data, a model matrix will be created for applying the regression model. The model matrix consists of two groups of features which will be generated.

- *Time based features*: Based on the time index separate information about quarter, month, day, hour and minute will be extracted. In addition, the time index will be converted to a numerical feature.
- *Imputed continuous data:* For the logistic regression model all previously prepared continuous features will be used as predictors.

Multiple Logistic Regression: For binary time series data, a multiple logistic regression will be used for anomaly detection.

$$\hat{P} = \frac{1}{1+exp^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots \beta_p x_p)}} \quad \text{Eq (2)}$$

$\hat{P} = \text{sigmoidal function of model terms with probability } [0,1]$
$x_p = \text{explanatory variable}$
$\beta_0 = \text{y\_intercept}$
$\beta_p = \text{slope coefficients for each explanatory variable}$

The dependent variable is one univariate binary feature and the independent variables are the features which has already been defined in the model matrix (x_p).

Impute anomalies: In the imputation process step all detected anomalies and NA will be replaced by predictions from the multiple logistic regression model.

This process starting from anomaly detection to anomaly imputation will be applied to every univariate binary time series. At the end all binary time series were prepared.

Output Data: After preparing continuous and binary data the following results are available:

- Partially imputed time series
- Fully imputed time series
- Detailed information about Monte Carlo simulation
- Detected anomalies

4. Missing data imputation using physics-based simulation models

Besides missing data imputation using data science techniques (see Section 2.3), two approaches using physics-based simulation models have been developed. These models use domain knowledge of DH networks or its components, which are encoded in simulation models. Simulation based models are useful for longer periods of missing data, where the underlying data generating process cannot be learned with sufficient accuracy from existing data, and for applications where physics-based model are well established and accurate. Data science techniques were applied to univariate time series, whereas simulation-based models are applied to multivariate time series, i.e. correlated measurements (e.g., temperature, pressure and mass flow rate). The first approach, called HT_sim (heat transfer station simulation), is bottom-up, whereas the second approach, called DHN_sim (district heating network simulation) is top-down. The innovative combination of the data science and the two simulated based approaches, as well as their differences, are described in Section 0.

4.1. Heat transfer station simulation HT_sim

4.1.1. Scope and applicability

The HT_sim method refers to heat transfer stations used for indirect connections to the DH networks via heat exchanger, which is common for all DH network topologies. As described in Section 2.1, the lowest common denominator of measurement equipment of a heat transfer station is the heat meter data used for billing (consumed energy). Heat meters typically consist of a volume flow sensor and inlet and outlet temperature measurements, the mass flow can be calculated from volume flow measurements. Depending on the measurement instrumentation and data logging, not all these data points are available. Data channels can be missing permanently, i.e. the respective sensors are not installed or data is not recorded. Oftentimes heat meter data from the primary side is available but only the flow temperatures of the secondary side, which are used for control purposes, as depicted in Figure 11.

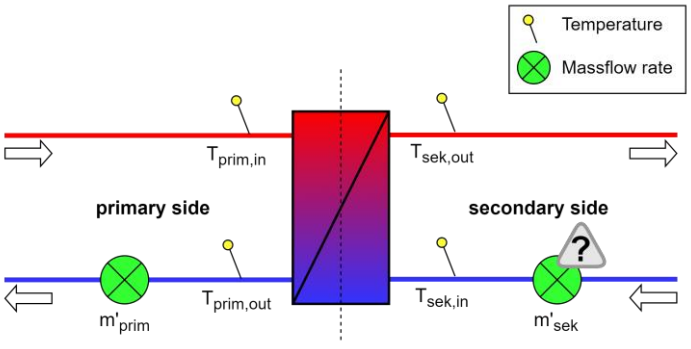


Figure 11: Measurement setup of a heat transfer station where the mass flow rate of the secondary side is not measured

In addition to permanently missing measurement data, temporarily data errors can occur, the instruments can be faulty or break down due to external causes. Errors can also occur during data transmission or

storage. If data is temporarily missing, the developed HT_sim method fits the simulation model to the characteristic working conditions of previous operating periods. If data is permanently missing, the underlying simulation models can also be applied, but in this case the heat exchanger specifications need to be known and factors like degradation to be manually accounted for [53]. An example of a heat transfer station where the primary-side return temperature and mass flow are temporarily missing is shown in Figure 12.

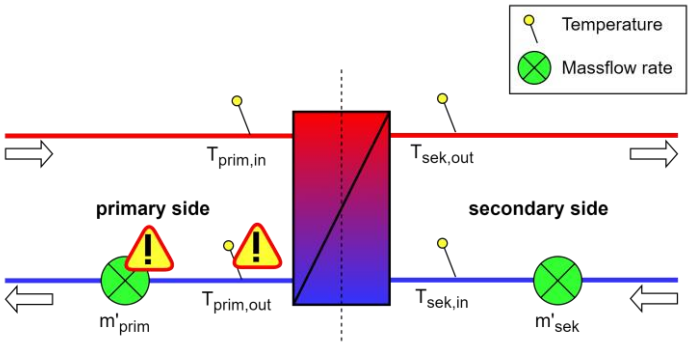


Figure 12: Measurement setup of heat transfer station with no return temperature and flow rate measurements on the primary side

The cases of missing data channels for which the HT_sim method can be used are categorized in Table 4. Note that the HT_sim method solves all cases where one input is missing (#1-#6) and 12 out of 15 cases where two inputs are missing (#7-#18). Not all cases can be solved, e.g., when both primary and secondary flow rates are missing. The procedure does not work if three or more inputs are missing.

*Table 4: Overview of cases that can be solved with the **HT_sim** approach. (X marks channels that are missing)*

Case	$\dot{m}_{prim}$	$T_{prim,in}$	$T_{prim,out}$	$\dot{m}_{sec}$	$T_{sec,in}$	$T_{sec,out}$	missing	modelName
#1	X						*1*	*PlateHEXepsNTU_m1*
#2		X					*1*	*PlateHEXepsNTU_T1in*
#3			X				*1*	*PlateHEXepsNTU_fieldTest*
#4				X			*1*	*PlateHEXepsNTU_m2*
#5					X		*1*	*PlateHEXepsNTU_T2in*
#6						X	*1*	*PlateHEXepsNTU_fieldTest*
#7	X	X					*2*	*PlateHEXepsNTU_m1T1in*
#8	X		X				*2*	*PlateHEXepsNTU_m1*
#9		X			X		*2*	*PlateHEXepsNTU_T1inT2in*
#10			X		X		*2*	*PlateHEXepsNTU_T2in*
#11				X	X		*2*	*PlateHEXepsNTU_m2T2in*
#12		X				X	*2*	*PlateHEXepsNTU_T1in*
#13			X			X	*2*	*PlateHEXepsNTU_fieldTest*
#14				X		X	*2*	*PlateHEXepsNTU_m2*
#15	X				X		*2*	*PlateHEXepsNTU_m1T2in*
#16		X		X			*2*	*PlateHEXepsNTU_m2T1in*
#17	X					X	*2*	*PlateHEXepsNTU_m1*
#18			X	X			*2*	*PlateHEXepsNTU_m2*

4.1.2. Step-by-step procedure

The step-by-step procedure for the HT_sim method is shown in Figure 13.

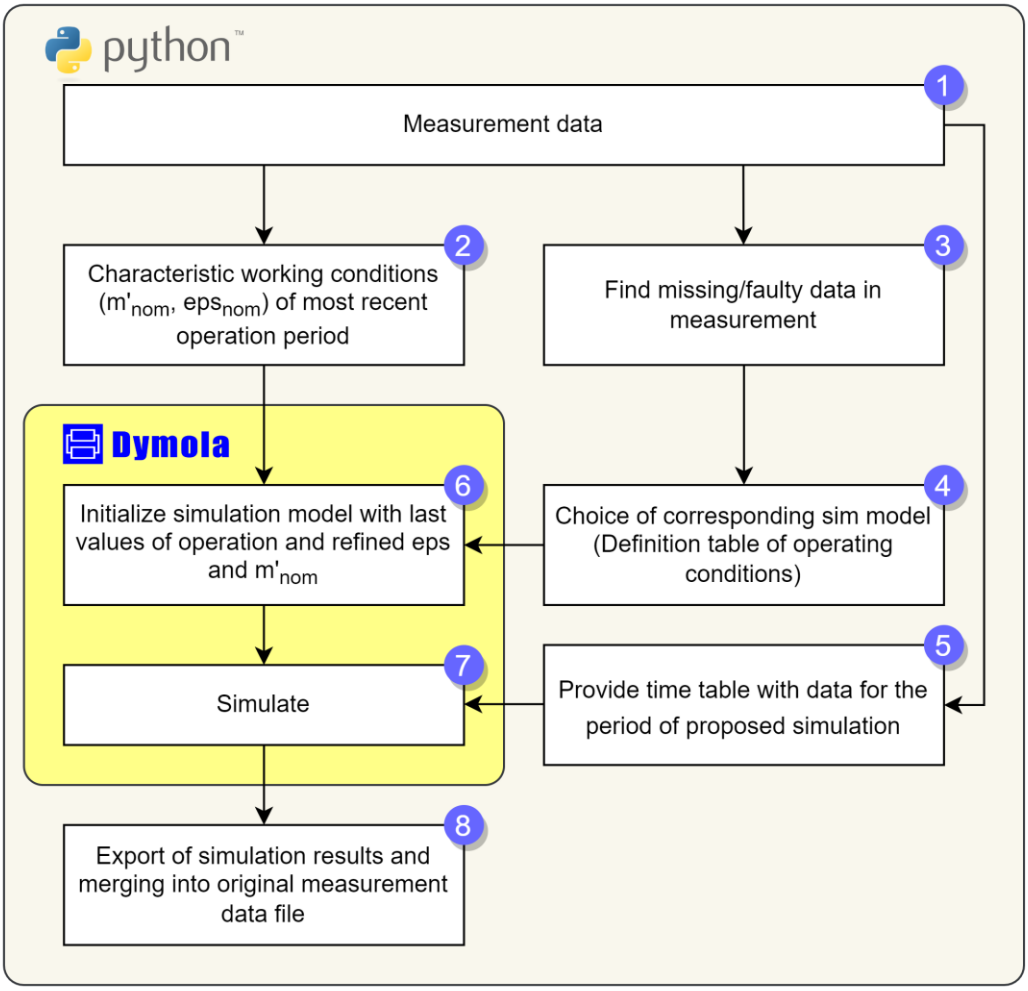


*Figure 13: Workflow of the **HT_sim** method*

Pre- and postprocessing steps (1-5, 8) are performed in Python. The initialization of the model and the simulation itself (6-7) is done in Dymola, invoked via call functions from within the Python script. The method entails the following steps:

1. Continuous analysis of measurement data from the last operational period.
2. Over a monitoring period of one month (backwards) the actual performance and operating values of the heat exchanger in the field are being analysed. Here the two different methods of calculating the heat exchanger effectiveness can be used (NTU-method and LMTD-method, for modelling background see [54])
3. In addition to the effectiveness, the nominal mass flows of the primary and secondary sides are determined. In this case the parameter estimation is done constantly with a moving observing

window of one month. This procedure can be improved by filtering historic data for stationary intervals.

4. The measurement data is checked for data gaps and validates if the method can be applied.
5. Based on the missing data channels, the appropriate simulation model is selected. (see Table 4)
6. A corresponding time series of the investigated time period with the remaining data channels available has to be provided.
7. Call Dymola (Modelica) from Python. Initialization of the chosen simulation model with last values of available measurement data (temperatures and flows) and parametrization of the heat exchanger model with the refined performance value (eps) from step (2).
8. Perform simulation.
9. Export simulation results. Merge results with the original measurement file.

4.1.3. Dymola simulation model

For the different missing data cases (see Table 4), simulation models were developed in Dymola. Some cases are redundant and the total number of models needed can be reduced to 10. An overview off all 10 simulation models can be found in Appendix A 8.2, and they are also available as code (see Appendix A 8.2). In the following, an example for case #8 in Table 4: Overview of cases that can be solved with the **HT_sim** approach. (X marks channels that are missing)Table 4 (missing data channels $\dot{m}_{prim}$, $T_{prim,out}$, see also Figure 12) is explained. The Dymola simulation template used for this case is shown in Figure 14.

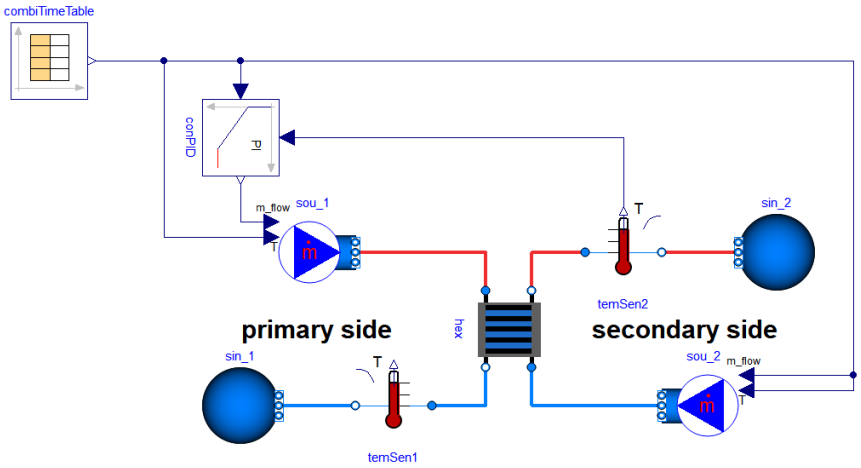


Figure 14: Scheme of the simulation template in Dymola for case #8

The heat exchanger, situated in the middle, is hydraulically connected to the boundaries of the primary and secondary side. The boundary of the secondary side is directly fed with values from the measurement data. The primary side, namely the primary side mass flow rate, is connected to a PI-controller. The control variable is the secondary side outlet temperature compared with the actual measurement values of the outlet temperature. The actuator variable is the primary side mass flow rate, regulating to reach the set

value of the temperature difference of 0 K. The same approach, using the remaining data to solve for the missing values, was used for the other cases.

4.1.4. Missing data imputation for use case

The following shows the application of the HT_sim method for measurement data of a real-world network Stanz DHN (Styria, Austria), for a detailed description see Section 4.2.1. The available data are from the 11 heat transfer stations in the network. Both the primary and secondary sides are measured. Measurement data of one customer for one week (2021-02-08 to 2021-02-14) is shown in Figure 15.

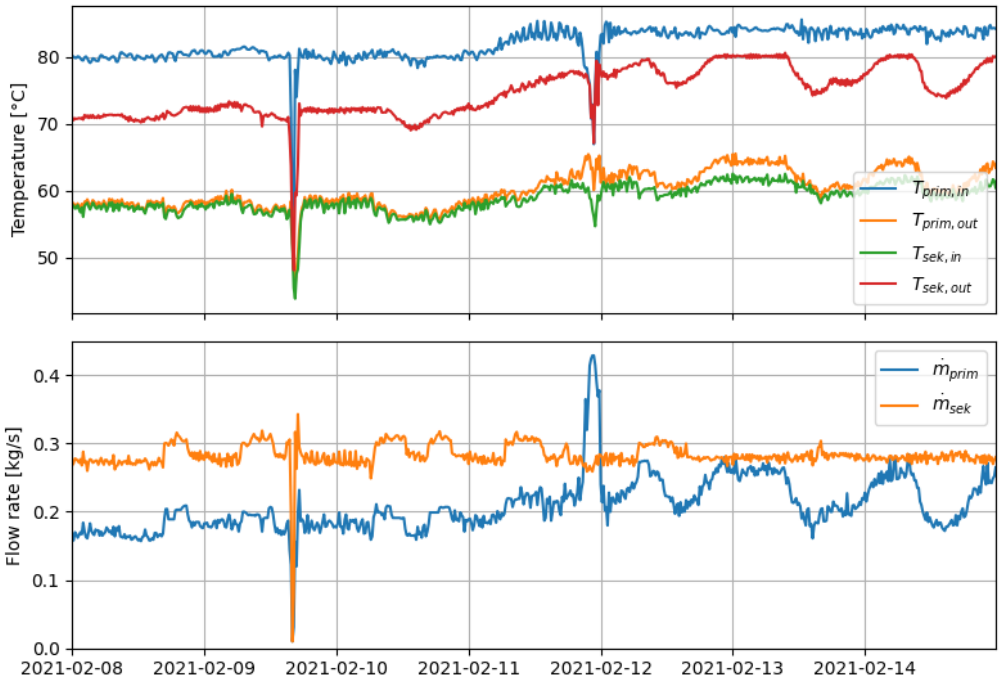


Figure 15: Measurement data of a heat transfer station in the Stanz district heating grid over the period of one week. Shown are the primary and secondary side temperatures (top) as well as the flow rates (bottom).

The presented results of the HT_sim method refer to case #8 as shown in the previous section (missing primary side mass flow rate and outlet temperature). For the purpose of demonstration, data gaps for $\dot{m}_{prim}$ and $T_{prim,out}$ were intentionally inserted to the measurement data for the period 2021-02-08 to 2021-02-14.

Following the workflow described in Section 0, the previous measurement period is first examined and the nominal values for parameterising the heat exchanger are determined from it. The specific values for this use case are shown in Table 5.

Table 5: Heat exchanger parameters derived from measurement data.

ϵ_{nom}	0.9468
$\dot{m}_{1,nom}$	0.2085 kg/s
$\dot{m}_{2,nom}$	0.2828 kg/s

These values are now used in the next step to parameterise the simulation model. The corresponding simulation model is determined using the matrix in Table 4. In this case the model *PlateHEXepsNTU_m1* is used. With the nominal parameters and the last correct measurement values of flows and temperatures the simulation model is initialized. The simulation time frame is congruent with the length of the respective data gap, in this show case one week.

Figure 16 shows the simulation results of the missing values of $T_{prim,out}$ and $\dot{m}_{prim}$ of the use case. In addition to the absolute values of the measurement and simulation, the relative errors of the respective variables are shown.

As shown, the approach delivers comparatively accurate values for longer periods of missing data. The relative errors on both, the inlet temperature as well as the flow rate, are within the range of +/- 2% during a regular operation period. However, the models fall short in moments with extremely high dynamics and times when the real-world operation is far outside the nominal working conditions.

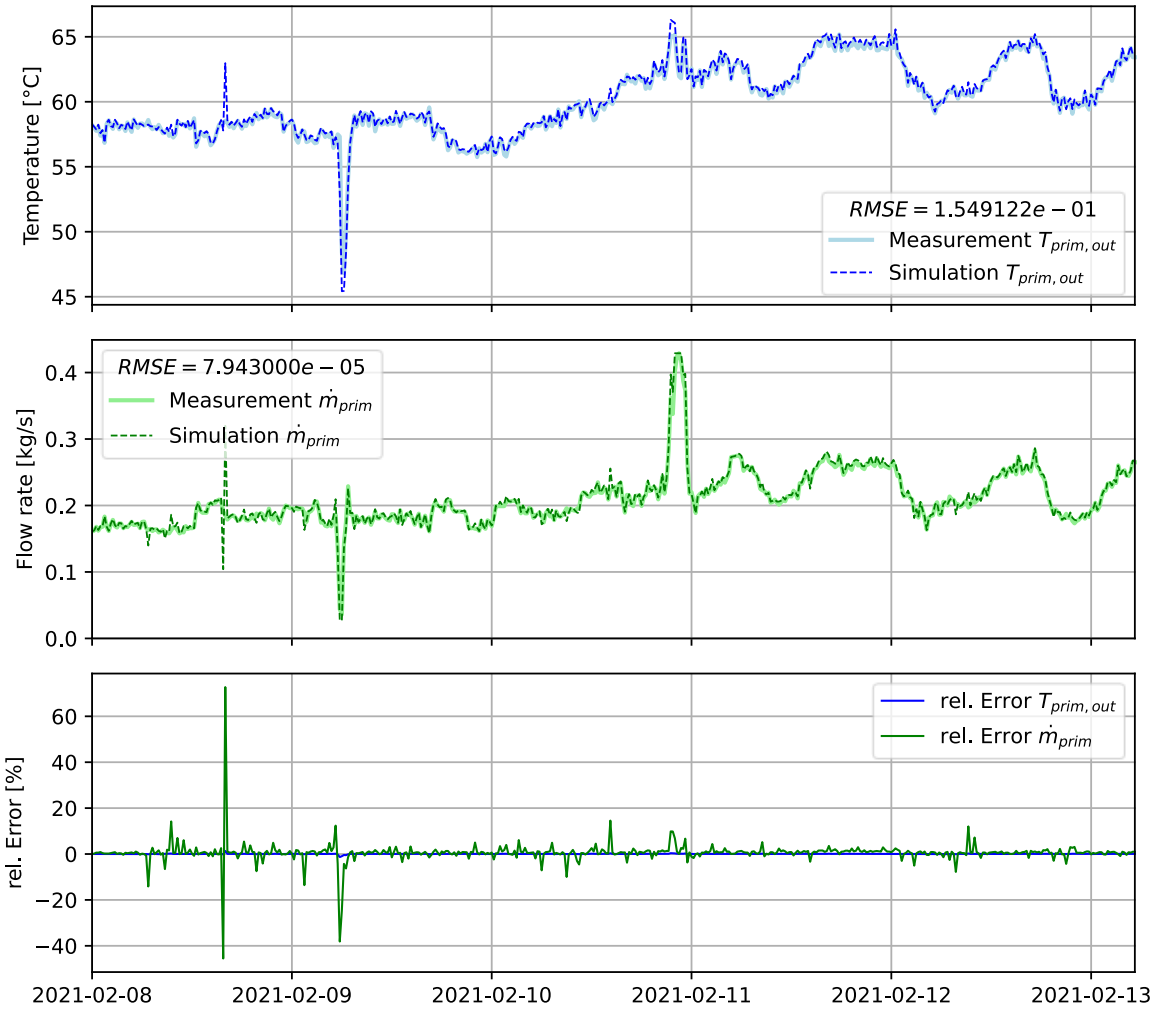


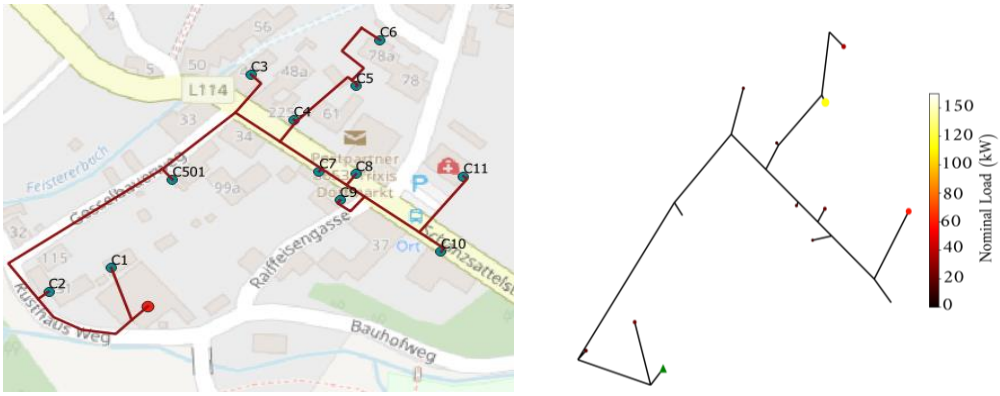
Figure 16: Comparison of simulation and measurement values for case #8 (primary side flow and outlet temperature missing) with absolute values and relative differences.

4.2. District heating network simulation DHN_sim

The DHN_sim method refers to missing data imputation based on a simulation model of the whole DH network. In this section, the method is applied and validated with the Stanz DHN. The simulation model is available as code (see Appendix A 8.1).

4.2.1. Description of Stanz DHN

Stanz DHN is a DH network in Stanz im Mürztal (Styria, Austria) with 630 m route length and 11 consumers. The topology of the network is represented on the left-hand side in Figure 17. Even though the consumers C501 and C10 are represented in the topology, they are not yet connected to the grid yet. The right-hand side in the figure shows the nominal loads of each consumer connected to the grid. The maximum nominal load belongs to the consumer C5 with 114 kW, the minimum load belongs to consumers C4 and C9 with 11 kW.



(a) network topolgy *(b) load distribrition*

Figure 17: Stanz DHN. On the left-hand side (a) topology of the network, on the right-hand side (b) distribution of the nominal load of each consumer in Python layout.

Measurement data are recorded from 2021-02-04 onwards in 1 min resolution. There are 179 features from the collected measurement data. 132 of them belong to the consumer side (11 features/ consumer). These 11 features include return and supply temperatures from both secondary and primary side of the network and mass flow rate at the primary side. On May 2021, due to the implementation of the new control strategies, the measurement data is missing for 1 month. The values of not connected consumers is assigned 0 instead of NA.

Figure 18 shows NA and 0 values in the data set between 2021-02-04 and 2022-08-01. As seen from the figure, C4 and C7 were connected to the network in the middle of October 2021. Additionally, the return temperature of the C7 on the secondary side is not available. In summer (starting from May 2021) due to the lack of heating demand, the mass flow rate of some consumers (e.g., C1, C6) contain 0 values.

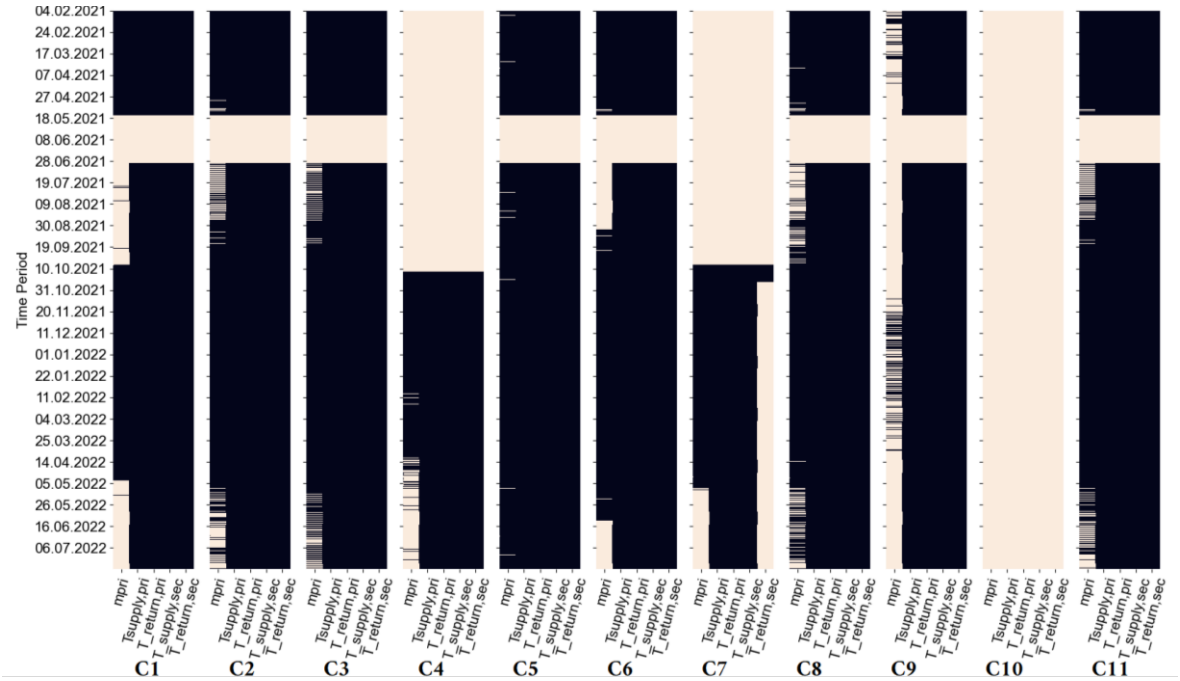


Figure 18: Heatmap of 0 and NA values (magenta) for 5 features in all consumers (C1 - C11)

Consumers C4 and C7 were connected to the network in October 2021, the measurement data used in the workflow is from 2021-10-14 till 2022-08-01 in 15 min resolution. The monthly average of the consumer load profile between this period is provided in Figure 19. This figure also supports the observation in Figure 18, that the highest share of 0 values in the consumer mass flow rate belong to C4 and C9.

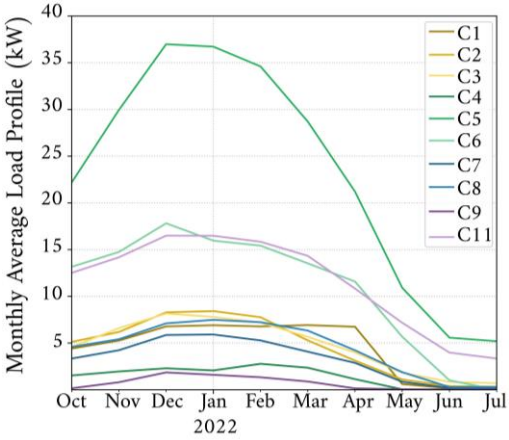


Figure 19: Monthly average of each consumer load profile between 2021-10 and 2022-07

4.2.2. Dymola simulation model

To set up the simulation model, the topology of an existing district heating network in shape file format is automatically translated to a structured model by NetworkX, a Python package for the creation and manipulation of complex networks by defining nodes and edges [55]. Each node in the structured model represents a DH consumer and production unit, and each edge represents a distribution pipe. The nodes and edges carry the information on the physical parameters of the network such as pipe lengths,

diameters, consumer load profiles in dictionary format. This package provides an error prone framework against the manual entries of the physical parameters of the pipes which can cause a wrong investigation.

After the network topology is defined in the Python environment, this representation is automatically translated into the Modelica environment for the simulation with the software Dymola as shown in Figure 20 [56]. Modelica is a standardized, acausal, equation-based modelling language used to model and simulate complex multidomain physical systems. Comparison studies of general-purpose tools and modelling paradigms for district-scale energy systems highlight that Modelica is a promising modelling language for DHC systems simulations since there are many open source and commercial libraries available containing components for DHC.

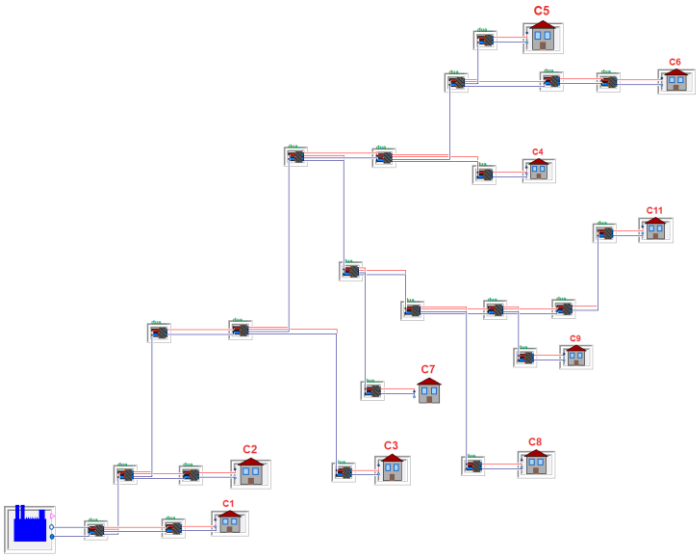


Figure 20: Top level of the Stanz DHN model in Dymola layout

A **consumer** consists of two pumps on the primary and secondary side, a heat exchanger and sink as shown in Figure 21. Nominal mass flow rate of the pump is calculated based on the nominal load in each consumer and assumed temperature difference. Variable return temperature is calculated based on the efficiency of the heat exchanger (HEX). The mass flow rate from the measurement data is assigned to the primary side of the pump. To prevent the numerical problems especially in the summer period, the smallest load mass flow rate assigned to the pump is kept at 0.0001 kg/s instead of 0 kg/s. Before the network simulation, a parameter optimization for the efficiency of each HEX is calculated.

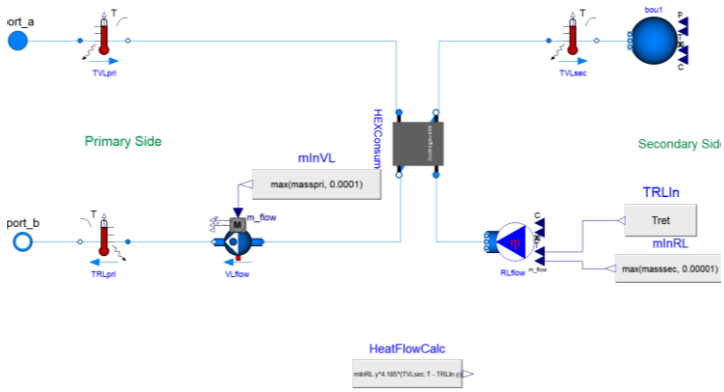


Figure 21: Consumer model

The **production unit** is modelled to maintain the minimum pressure difference at each consumer in the network by a PI-Controller [57] as shown in Figure 22. The model monitors the differential pressures of all sub-stations. A PI-controller controls the pressure of the supply pipe based on the minimum differential pressure in the entire network since the critical node could vary during the operation. In the production unit models, the fluid passing through an ideal heater is heated up to a set temperature. This set temperature is taken from the measurement data.

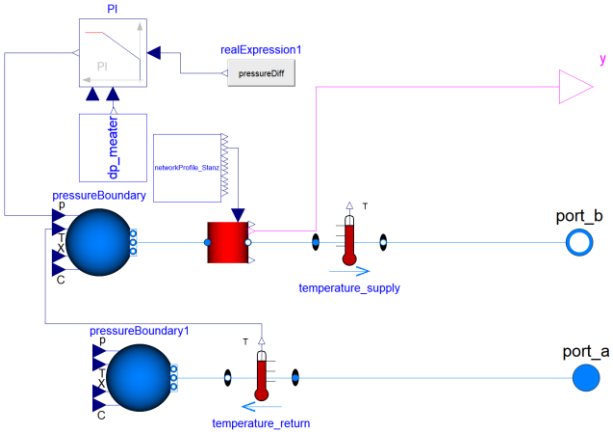


Figure 22: Production unit model

Pipes used in the Stanz DHN are twin pipes thanks to the latest improvements in pipe design and insulation materials. Twin pipes include two media pipes (return and supply) in the same casing which is a polyurethane foam represented in Figure 23. In a typical case of the twin pipe is that the two single pipes are identical, placed horizontally or vertically and in the same depth from the ground surface. The Dymola model is shown in Figure 24.

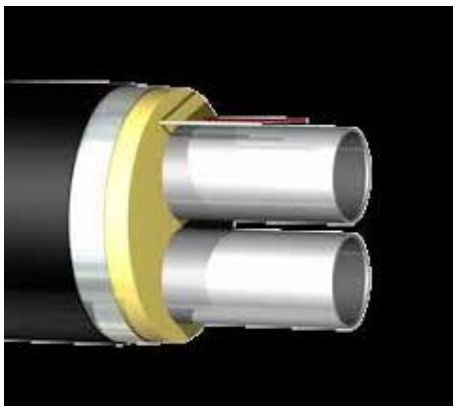


Figure 23: Twin pipe representation (source: Logstor)

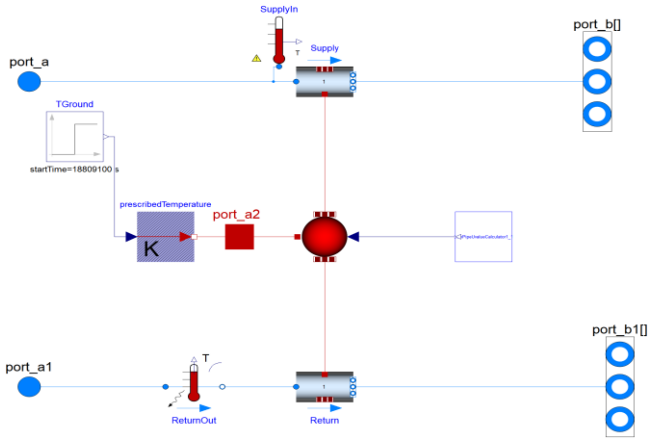


Figure 24: Pipe model

Since the return and supply pipes are thermally coupled with each other in addition to the ground, the heat losses based on this coupling is given in Eq (3) and Eq (4) [58].

$$q_1 = (U_{11} - U_{12})(T_1 - T_g) + U_{12}(T_1 - T_2) \quad \text{Eq (3)}$$

$$q_2 = (U_{22} - U_{21})(T_2 - T_g) + U_{21}(T_1 - T_2) \quad \text{Eq (4)}$$

The U values are the quantity often supplied by the pipe manufacturers. However, in the case of Stanz DHN where Kingspan-Logstor twin pipe serie 3 is used, U values were not available. Therefore, the U values are calculated based on the formulas and principles in accordance with the Wallenten et al. [59]. In this approach, the heat loss from each pipe is split into heat loss from the supply pipe to the return pipe, referred as asymmetric and heat loss from the entire pipe to the ground, referred as symmetric. The calculation is included in the Modelica pipe model where the heat transfer coefficients $h_{symmetric}$ and $h_{asymmetric}$ is calculated based on the First-Order Approximation.

$$U_{11} = U_{22} = (h_s + h_a)2\pi\lambda_{ins} \quad \text{Eq (5)}$$

$$U_{12} = U_{21} = (h_a - h_s)2\pi\lambda_{ins} \quad \text{Eq (6)}$$

For the heat losses the parameter values used in the simulations are given in Table 6.

Table 6: Parameter values for pipe heat losses applied in simulation

Parameter name	*Value*
Insulation (polyurethane foam) thermal conductivity [60]	0.023 (W/mK)
Ground thermal conductivity [58]	1.5 (W/mK)
Ground temperature	5/12 °C

4.2.3. Model validation

Table 7 shows the inputs and outputs in the Stanz DHN simulation. The inputs refer to the measurement data used as inputs in the simulation.

Table 7: Classification of the inputs and outputs in the network simulation

	Inputs	*Outputs*
Producer	• Supply Temperature Profile	• Pressure drops • Return temperature profile • Heat flow rate
Consumer	• Mass flow rate at the primary and secondary side • Secondary side return temperature profile • HEX efficiency (calculated)	• Primary return temperature profile • Secondary supply temperature profile
Pipe	• Physical parameters (length, diameters, casing material)	• Heat losses

For the network validation, first the error caused by the pipe model is considered. The supply temperature is decreased from the production unit to the consumer due to heat losses. Figure 25 shows the difference of the measurement and simulation of the inlet temperature profile of each consumer. The difference is provided in monthly average in error box plots where the median of each month is represented in red lines. The y-axis is kept between 10 °C to -10°C. Negative values mean that the simulation values are larger than the measurement values and heat losses are smaller than the expected. The y-axis shows the monthly average mass flow rate of each consumer which are plotted in the gray bar plots. In summer, the difference between measured and simulated values are higher compared to other months. The median is negative. Due to the lack of heating demand in summer, the mass flow rates are smaller than 0.01 kg/s. The implemented twin pipe model has a limitation when the mass flow rates are lower than 0.01 kg/s. As mentioned before, the consumers C5 and C11 have the highest load profile where the average monthly

mass flow rates are also higher and the difference of the temperature values in the pipe model due to the seasonality are minimized.

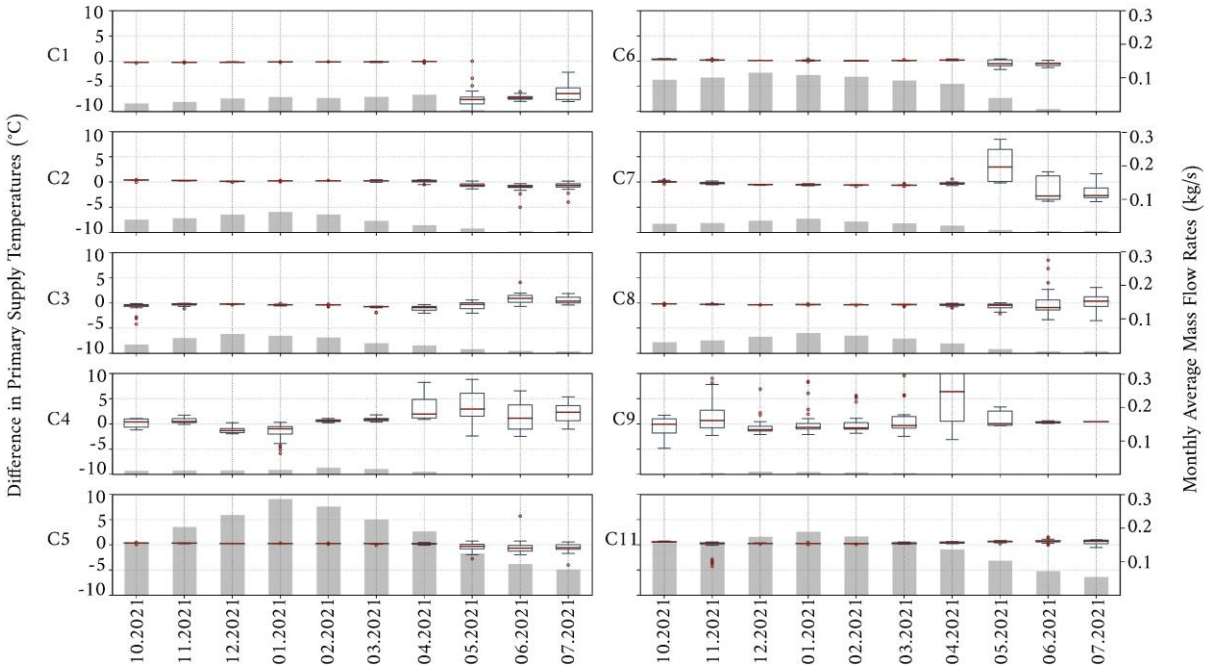


Figure 25: Difference of the measurement and the simulation (primary side) supply temperature in each consumer

Figure 26 shows the return temperature profile at the production units and the difference (measurement-simulation) in the return temperature also includes the error from the HEX in each consumer. The median of the difference aligns around 0 °C for the months between October 2021 and April 2022 where the RMSE is less than 1. However, the months between May and the end of July it increases to 3.55 due to the lower mass flow rates.

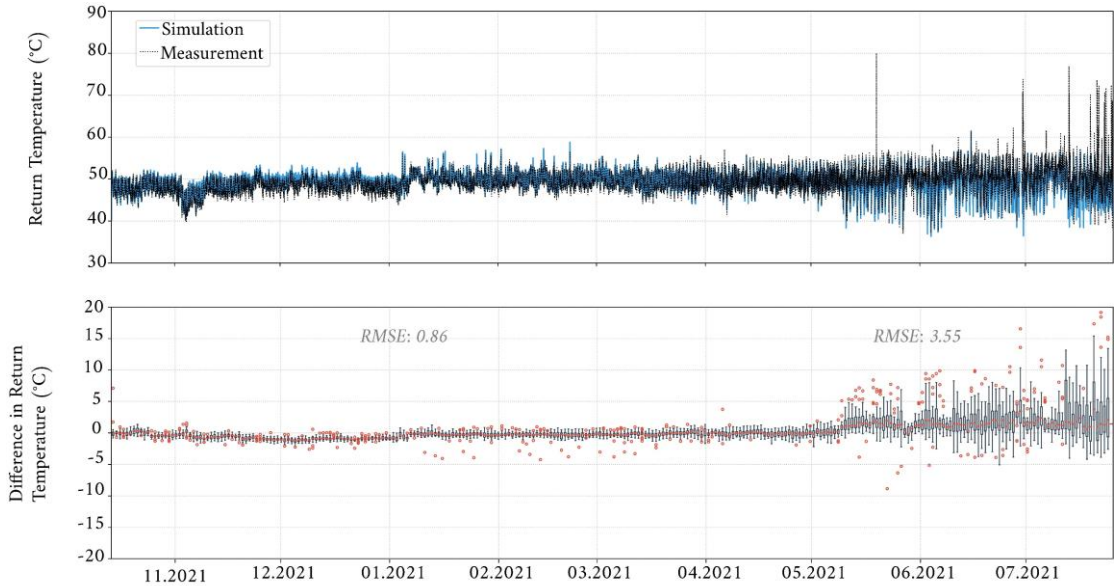


Figure 26: Comparison of the measurement and the simulated return temperature to the production unit

Figure 27 shows the load profile of the consumer. The demand is calculated on the secondary side in order to see the errors from the heat exchanger. RMSE of the load for each consumer varies between 0.2 to 1 except C5 (2.8) an C11 (1.65) which are the consumers with the highest heating demand. Since the mass flow rates are the same, the error stems from the temperature difference. The difference of the heating demand of these consumers varies between +/-2 kW.

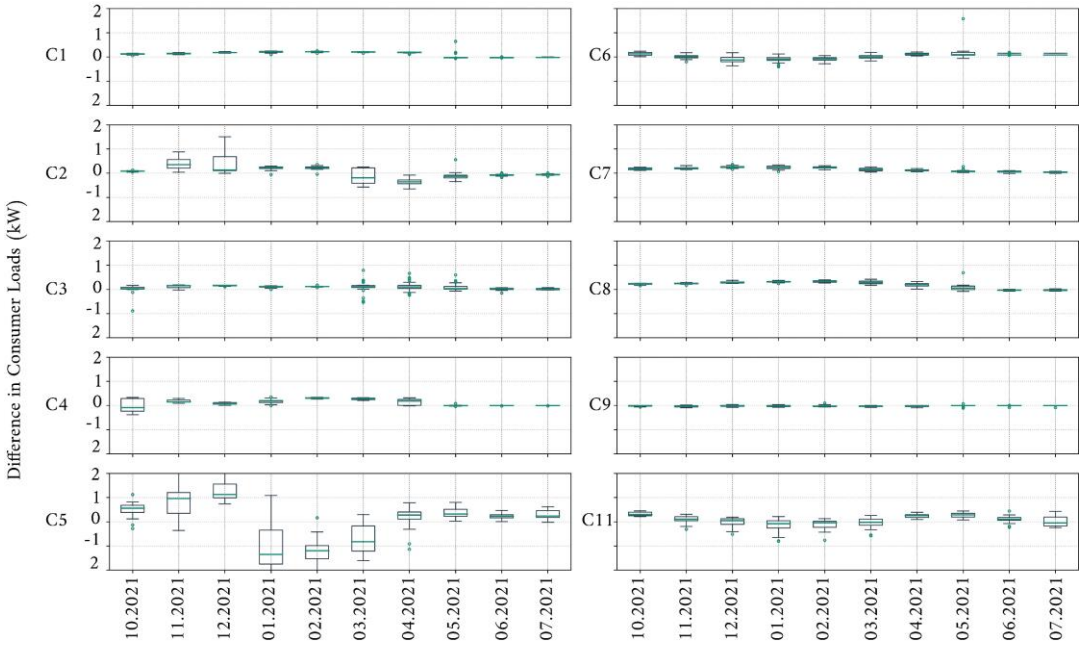


Figure 27: Difference of the measurement and the simulation load in each consumer

4.2.4. Missing data imputation for use case

In this workflow, the mass flow rate at consumer C11 on the primary side is assumed to be missing. The usual network model that uses the consumer and production unit model as described in Section 4.2.2 is referred as *Model A*. To calculate the mass flow rate in C11 with the simulation, these models are required to be modified since the unknowns and I/O profiles will be different, leading to *Model B*. The information of the consumer known and unknowns, inputs and outputs are given in the Table 8.

Table 8: Known/Unknowns for Model A and Model B for the consumer model

	$T_{supply,primary}$	$T_{return,primary}$	$T_{supply,secondary}$	$T_{return,secondary}$	$m_{primary}$	$m_{secondary}$
Model A	✓ Pipe carries this information	× Output: Calculated at the exit of HEX	× Output: Calculated the exit of HEX	✓ Input: Assign measurement data	✓ Input: Assign measurement data	✓ Input: Assign calculated profile based on $m_{primary}$
Model B	✓ Pipe carries this information	× Output: Calculated at the exit of HEX	× Output: Calculated the exit of HEX	✓ Input: Assign measurement data	× **Output: Calculated in the model (missing)**	× Input: Assign an estimated value

The consumer model in *Model A* takes the mass flow rate on the primary and secondary side as an input; the consumer model in *Model B* has to calculate these mass flow rates. However, since we do not know the primary side of the mass flow rate, we cannot calculate the secondary side mass flow rate. Therefore, we need to make an assumption on the secondary side mass flow rate when the primary side mass flow rate is missing. Eq (7) shows the assumption of the secondary side of the mass flow rate. The secondary side supply temperature is assumed as the $(T_{supply,primary} - 8)$ °C.

$$m_{secondary} = \frac{(T_{supply,primary} - T_{return,primary}) * m_{primary}}{(T_{supply,primary} - 8 - T_{return,secondary})} \quad \text{Eq (7)}$$

Beside the consumer model, the production unit also needs to be modified. The production unit in *Model A* calculates the mass flow rate based on the minimum pressure drop in the network; the production unit in *Model B* requires the mass flow rate profile as an input in the model.

The switch between Model A (as described in Section 4.2.2) Model B (modified for missing mass flow rate on primary side) is as follows: Model A is run until a defined stop time for periods where the primary mass flow rate is available. Model B is initialized with the Model A results and the simulation continued for periods when data is missing. When data is available again, Model A is initialized with the Model B results. This process is continued iteratively. The .mos scripts in Figure 28 shows the workflow.

```
modelA="EnableDHNModelica.Models.Stanz2022_StatusQuo_MassFlow";
modelB=" EnableDHNModelica.Models.Stanz2022_StatusQuo_MassFlow_enb";

StopTime1 = 2592000 // 30th day
StopTime2 = 6480000 // 75th day
StopTime3 = 17280000 // 200th day

// Simulate ModelA
translateModel(modelA);
simulateModel(modelA, startTime=0, stopTime=StopTime1-900,
numberOfIntervals=0, outputInterval= 900, resultFile="StaticA1");
system("copy dsfinal.txt dsfinal_Static.txt")
// Simulate ModelB
importInitial("dsfinal_Static.txt");
simulateExtendedModel(modelB, startTime = StopTime1-900, stopTime =
StopTime2, outputInterval = 900, resultFile="StaticB");
system("copy dsfinal.txt dsfinal_Missing.txt")

// Simulate ModelA Again
importInitial("dsfinal_Missing.txt");
simulateExtendedModel(modelA, startTime = StopTime2-900, stopTime =
StopTime3, outputInterval = 900, resultFile="StaticA2");
system("copy dsfinal.txt dsfinal_Continue.txt")
```

Figure 28: Worklow of DHN_sim method, switch between Model A and B

Figure 29 shows comparative evaluations with Model A and Model B for consumer C11. The top subplot shows the mass flow rate for Model A (black) and Model B (red). The bottom left subplot shows the error box of the average daily difference of the measurement and the simulation mass flow rates for the imputation period (25th day – 75th day). The median of the difference varies around -0.01 kg/s. A negative sign implies the mass flow rate from the Model B is assumed higher than the measurement data. The bottom right subplot shows the heat demand of Model A and Model B for the primary (black) and the secondary side (gray). The RMSE on the primary side of the demand is smaller than the secondary because the mass flow rate of the secondary side had to be assumed due to the missing mass flow rate at the primary side. Overall, the introduced errors for Model B are small, which shows the practicability of the method. The significant downsides identified with this method are the need to have a full, validated simulation model of the network, and that fact that for many types of sensor, only one of that type of sensor can be missing on the network in any given time period for the method to work.

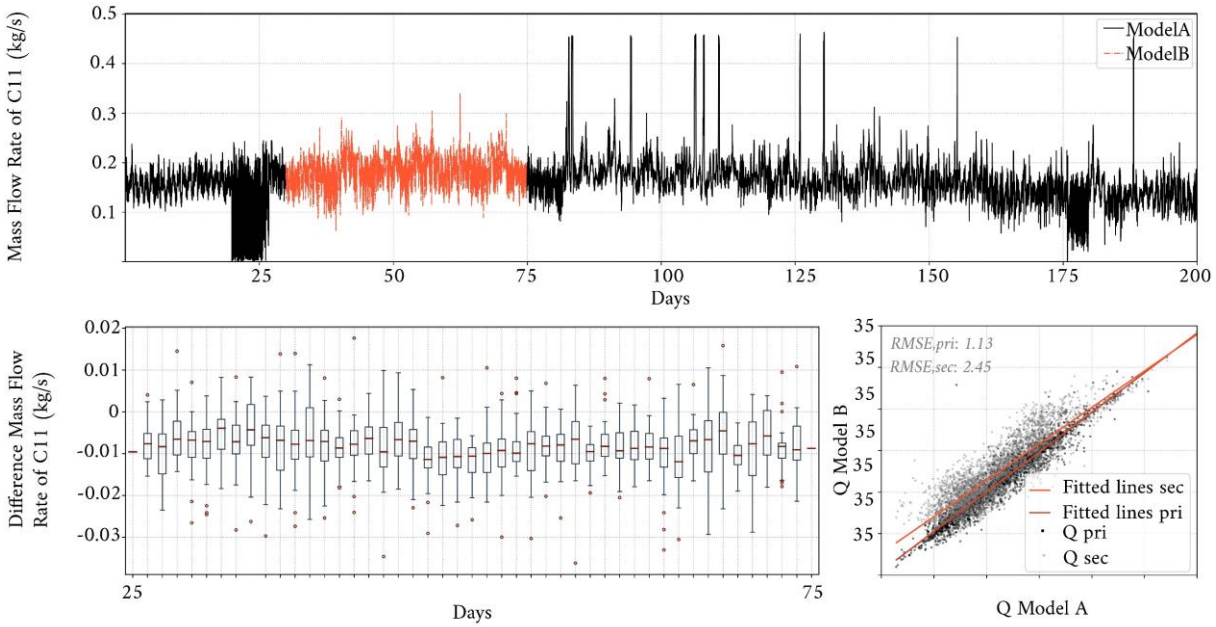


Figure 29: Results from simulation-based imputation. Mass flow rate at the missing consumer C11, average daily difference of the real and simulation mass flow rates

5. Combination of missing data imputation techniques

In Section 3 and 4, three missing data imputation methods were presented. To make best use of these methods, the following combined approaches have been conceptualized (Table 9).

Table 9: Combined missing data imputation techniques (DPP = Data pre-processing pipeline, HT_sim = Heat transfer station simulation, DHN = District heating network simulation)

Abbreviation	Methods
DPP_HT	DPP + HT_sim
DP_HT_DHN	DPP + HT_sim + DHN_sim

Before the combination of these methods is discussed, it is important to clarify their characteristics, strengths and weaknesses. Figure 30 provides an intuitive understanding, whereas Table 10: Comparison of DPP, HT_sim and DHN_sim methodsTable 10 compares the methods systematically.

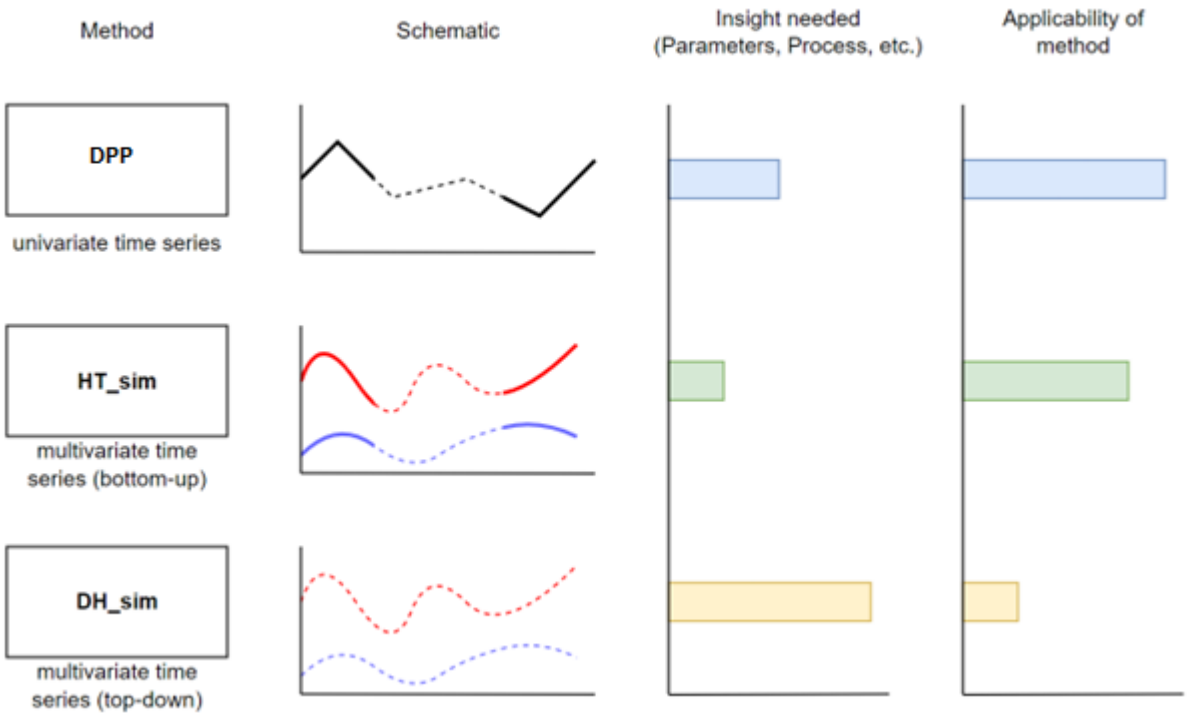


Figure 30: Comparison of missing data imputation methods

Table 10: Comparison of DPP, HT_sim and DHN_sim methods

	DPP Data pre-processing pipeline	**HT_sim** Heat transfer station simulation	**DH_sim** District heating network simulation
Method cluster	Data science / machine learning based	Physics-based simulation	Physics-based simulation
Scope	Anomaly detection and missing data imputation for binary and continuous univariate time series	Missing data imputation for multivariate time series for heat transfer station data	Missing data imputation of multivariate time series for district heating networks
Approach	bottom-up (single data channel)	bottom-up (component of DH network)	top-down (whole DH network)
Maximum allowed data gap	Depending on data channel, typically up to 25% of the number of data points within a day	One week or longer, depending on boundary conditions	One week or longer, depending on boundary conditions
Calculation speed	very fast (seconds)	fast (seconds to minutes)	medium to slow (minutes), depending on model complexity
Strength	- Highly robust, stable and generally applicable process, covering all typical DH network data channels	- Detailed and precise method - Performance values used in simulation are derived from most recent operating conditions, degradation and fouling of heat exchanger is accounted for - No information of heat exchanger needed, performance value are automatically estimated from measurement data - Applicable to any size of heat exchanger, adaptability to different heat exchanger designs (shell/tube-HEX, plate-HEX) - Filling of longer data gaps possible - Filling of permanently missing data channels possible	- Applicable when both primary and secondary mass flow of heat transfer station are missing (not covered by HT_sim) - Fully automated workflow even for complex DH simulation models - Filling of longer data gaps possible - Filling of permanently missing data channels possible
Weakness	- Data imputation for longer periods not possible	- Does not cover all missing data cases for heat transfer stations, maximum of two channels missing - Unreliable for operating conditions with extremely high dynamic or outside the nominal working range	- Data of only one consumer in the grid can be missing, data of other consumer needs to be known - Simulation model for the complete network needs to be set up

For any innovative combination of these methods, DPP should always be applied first to detect anomalies in the data, as HT_sim and DHN_sim do not include anomaly detection, but are limited to data imputation.

In the first step, DPP will eliminate statistical outliers and significant measurement errors (physically implausible values), duplicates, error codes, strings, etc. In the second step, DPP does data imputation for the detected anomalies and missing values. As the imputation does not include explicit domain knowledge, it is crucial to limit the maximum gap size, as otherwise huge missing data gaps could lead to distorted results and inaccuracies. Based on analyses of DH data, a good general rule is to limit missing data to 25% of the number of data point within a day. DPP could be fine-tuned depending on the DH network and the data channel, where data channels with higher dynamics (e.g.; thermal power) should have shorter maximum gaps than data channels with less dynamics (e.g.; ambient temperature). DPP can be used as a standalone tool, and it is recommended to use DPP for any subsequent analysis or simulation-based missing data imputation.

DPP_HT combines DPP and HT_sim in sequential order. The main advantage of this approach is, that it is straightforward to set up the pipeline for any DH network as it needs only component level models for heat transfer stations, which are very similar across the globe. HT_sim does not require parametrization as its performance values are derived from measurement data and is able to automatically select the appropriate model depending on which data channels are missing. This innovative combination has therefore the potential to be fully automated for DH networks of arbitrary size. Compared to using DPP as a standalone tool, the combination of DPP and HT_sim is able to fill longer missing data gaps and data channels which are permanently missing. However, data imputation for longer gaps or permanently missing data channels is limited to heat transfer stations (and two missing data channels per heat transfer station).

DPP_HT_DHN combines DPP, HT_sim and DHN_sim in sequential order. The main advantage of this approach is, that it covers additional missing data cases compared to DPP_HT, i.e. when both the primary and secondary mass flow rate at the heat transfer station are missing. The main disadvantage of this approach is, that a simulation model for the complete network needs to be set up and only a very peculiar case is addressed, namely that data of exactly one consumer in the grid is missing. Unless a network simulation model is already available, the additional benefit does not seem to justify the additional effort. Also, the potential to be fully automated is not given, as each network simulation needs considerable manual fine-tuning. Therefore, this combination is unlikely to be of widespread practical use. The same is true for the combination of DPP and DHN_sim (without HT_sim).

6. Discussion, conclusion and outlook

As shown in this report, data-driven analytics and machine learning methods are able to address major challenges in the district sector and are already extensively used for applications like load forecasting, weather forecasting, fault detection, predictive maintenance and control and optimal scheduling. The value created by data-intensive techniques crucially depends on the quantity and quality of the available data.

An analysis of the state-of-the art of ML applications to DH networks (see Section 2.2) showed, that the deployed ML methods and what are they used for is well researched. ML is mostly used in load forecasting, since the prediction of building energy usage plays a vital role in developing a model predictive controller for consumers and optimizing the energy distribution plan for utilities. ANN, regression based and SVM approaches are the most used methods overall. These methods are well established and commonly used in the ML field, which reiterates that in terms of the used ML methods, the district heating sector is not lagging behind.

However, the analysis showed, that data requirements have been largely neglected in the scientific literature and no comprehensive review exists for data requirements of ML methods in the context of DH. A first-hand analysis of 63 papers showed, that about half of the publications report data problems and almost all of these papers explain the data processing in some detail. About half of the publications do not report data problems and do not explain the data processing in detail. As data scaling and other data manipulation is necessary for most ML method to improve prediction performance, it can be assumed that some data processing is performed even if the data quality would be perfect. For papers who do report on data processing, the deployed methods for outlier detection and gap filling are largely practical heuristics (e.g.; 60% of papers do simple linear interpolation for gap filling), rather than applying state-of-the-art data science methods. To make research results more traceable, a standardized reporting format and guideline for data processing would be highly desirable. Ideally, data processing should be made with Open Source tools where the deployed methods are available and reproducible for the community.

As shown in Section 2.1, data acquisition, transmission and storage deployed in DH networks has a low degree of standardization and depends on the age and degree of modernization of the network and various other parameters. The common denominator for the recorded data channel is, that DH data consist of binary and continuous time series and that the recorded measurement channels of certain components, in particular heat transfer stations, can be reduced to heat meter recordings (mass flow rate / volume flow rate, inlet temperature, outlet temperature). For a generic framework with the potential to be fully automated, it is therefore advisable to focus on data science methods which can deal with all binary and continuous univariate time series for typical DH data and use physics-based simulation models of for highly standardized components.

As shown Section 2.3, a major data mining challenge for DH time series is, that these are not necessarily statistically independent from their past (or future). For anomaly detection, dedicated methods for these

time series have recently been shown to be unreliable, therefore the decision was made to stick with methods that were primarily designed for datasets consisting of independent and identically distributed data. Pre-processing of DH time series requires that several deterministic components need to be extracted from the raw data, e.g., trends, changepoints, and seasonality. However, it must be noted that the state-of-art in this sub-area of data science is less explored than forecasting. Hence, in practice it is often necessary to design application-specific solutions that are tailored to the time series data at hand. Time series data imputation as a scientific discipline is not as mature as one might expect, developments in the field should be incorporated to improve data pre-processing pipelines.

The developed data pre-processing pipeline (DPP), as described in Section 3, builds on highly robust and stable state-of-the art of data science methods. Compared to widespread methods among practitioners in the district heating field, DPP is a substantial improvement, as makes a clear methodological distinction between continuous and binary data, incorporates the season length, builds on state-of-art anomaly detection method (Isolation Forest, Robust Kernel Density Estimation, Subspace Outlier Detection, Dirichlet Process Mixture Models) and offers a variety of missing data imputation methods (interpolation, moving average and season decomposition methods) which are selected according to a Monte Carlo simulation. A main advantage of this method is, that it is generally applicable and can be fully automated. It can be applied as a standalone tool independently of the subsequent methods. To further spread its usage, future research project should develop the tool further from the current proof-of-concept status, fine-tune it to specific boundary conditions where necessary, show its practical application based on public data sets and establish an appropriate governance structure to secure long-term maintenance as an Open Source project.

Although about a quarter of the current research papers uses simulation data, no application could be found where measurement and simulation data were combined in an integrated workflow to facilitate ML applications. Two new physics-based approaches, as described in Section 4, were developed, namely the bottom-up approach “heat transfer station simulation” (HT_sim) and the top-down approach “district heating network simulation” (DHN_sim). Compared to the DPP approach, these models allow to fill longer missing data gaps (up to one week or even longer) and data channels which are permanently missing as they include domain knowledge from heat transfer stations and district heating network respectively. A main advantage of the HT_sim approach is, that this process is straightforward to set up for any DH network, it only needs component level models for heat transfer stations, which are very similar across the globe. Also, HT_sim choses the appropriate model automatically depending on which data channels are missing.

The main advantage of DHN_sim compared to HT_sim is that it covers additional missing data cases, i.e., when both the primary and secondary mass flow rate at the heat transfer station are missing. However, DHN_sim requires a complete network simulation model and only covers a very peculiar missing data case, namely that data of exactly one consumer in the grid is missing. This additional benefit hardly justifies its use in practice.

Among other, the combination of DPP and HT_sim in sequential order in the innovative concept DPP_HT was analyzed in the project (see Section 5). Compared to using DPP as a standalone tool, the combination of DPP and HT_sim is able to fill longer missing data gaps and data channels which are permanently missing. This approach has a high potential for the district heating community and future research projects should work towards fully automated workflows for DH networks of arbitrary size. Regarding the HT_sim method, future research should focus on specific improvements like the handling of highly dynamic operating conditions, improve model parametrization and initialization and possibly extend the model to cases with more than two missing data channels. The extension of DPP_HT to additional system components and production units that have a high degree of standardization in terms of typically available data channels and modeling properties, e.g., biomass boilers, oil boilers, gas boilers, solar thermal plants, heat pumps and water storages, should be taken up in future projects to give this promising method a broader scope. The implemented procedures should be available as Open Source libraries to make data processing as transparent as possible. Full-scale DH network simulations for missing data imputation on the other hand is not regarded as a promising direction for future research.

Of the analyzed studies of ML applications in the DH sector, only 17% used public data sets, whereas 83% used private datasets. In terms of open data, the energy sector ostensibly lags behind other fields [13]. To make research work more traceable and improve the collaboration on data-related issues, the district heating community should strive towards scientific data management and stewardship as outlined by the FAIR Guiding Principles [61] or similar. Future research projects should further elaborate the effect of data quality on the prediction accuracy of ML models, which can be done on a large scale only with public data sets.

7. References

[1] “Renewables 2021 Global Status Report,” REN21 Secretariat, Paris, 2021.
[2] B. Möller, E. Wiechers, U. Persson, L. Grundahl, R. S. Lund, and B. V. Mathiesen, “Heat Roadmap Europe: Towards EU-Wide, local heat supply strategies,” Energy, vol. 177, pp. 554–564, Jun. 2019, doi: 10.1016/j.energy.2019.04.098.
[3] H. Lund et al., “4th Generation District Heating (4GDH),” Energy, vol. 68, pp. 1–11, Apr. 2014, doi: 10.1016/j.energy.2014.02.089.
[4] M. A. Sayegh et al., “Trends of European research and development in district heating technologies,” Renewable and Sustainable Energy Reviews, vol. 68, pp. 1183–1192, Feb. 2017, doi: 10.1016/j.rser.2016.02.023.
[5] H. Lund et al., “Perspectives on fourth and fifth generation district heating,” Energy, vol. 227, p. 120520, Jul. 2021, doi: 10.1016/j.energy.2021.120520.
[6] A. R. Mazhar, S. Liu, and A. Shukla, “A state of art review on the district heating systems,” Renewable and Sustainable Energy Reviews, vol. 96, pp. 420–439, Nov. 2018, doi: 10.1016/j.rser.2018.08.005.
[7] L. Lyons, “Digitalisation: Opportunities for heating and cooling.” Publications Office of the European Union, 2019.
[8] C. Ntakolia, A. Anagnostis, S. Moustakidis, and N. Karcanias, “Machine learning applied on the district heating and cooling sector: a review,” Energy Syst, Jan. 2021, doi: 10.1007/s12667-020-00405-9.
[9] G. Mbiydzenyuy, S. Nowaczyk, H. Knutsson, D. Vanhoudt, J. Brage, and E. Calikus, “Opportunities for Machine Learning in District Heating,” Applied Sciences, vol. 11, no. 13, p. 6112, Jun. 2021, doi: 10.3390/app11136112.
[10] “DIRECTIVE (EU) 2018/ 2002 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL - of 11 December 2018 - amending Directive 2012/ 27/ EU on energy efficiency,” p. 21.
[11] Q. Sun et al., “A Comprehensive Review of Smart Energy Meters in Intelligent Energy Networks,” IEEE Internet Things J., vol. 3, no. 4, pp. 464–479, Aug. 2016, doi: 10.1109/JIOT.2015.2512325.
[12] DHC+ Technology Platform, “Digital Roadmap for District Heating & Cooling,” Euroheat and Power, no. July, pp. 1–40, 2019.
[13] S. Pfenninger, J. DeCarolis, L. Hirth, S. Quoilin, and I. Staffell, “The importance of open data and software: Is energy research lagging behind?,” Energy Policy, vol. 101, pp. 211–215, Feb. 2017, doi: 10.1016/j.enpol.2016.11.046.
[14] F. C. L. Trindade, L. F. Ochoa, and W. Freitas, “Data analytics in smart distribution networks: Applications and challenges,” in 2016 IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia), Melbourne, Australia, Nov. 2016, pp. 574–579. doi: 10.1109/ISGT-Asia.2016.7796448.
[15] J. Westerweck, “Insights from digitalization of a supplier – Digitalisation in District Heating, Vattenfall Wärme Hamburg Gmbh,” presented at the IEA DHC Annex TS4 - Workshop, Frankfurt.
[16] R. Wiltshire, “Advanced District Heating and Cooling (DHC) Systems”.
[17] M. P. Deisenroth, Mathematics for Machine Learning, 1st ed. Cambridge ; New York, NY: Cambridge University Press, 2020.
[18] E. Alpaydin, Introduction to Machine Learning, fourth edition, Fourth edition. Cambridge, Massachusetts: The MIT Press, 2020.
[19] K. P. Murphy, Probabilistic Machine Learning: An Introduction. Cambridge, Massachusetts: The MIT Press, 2022.
[20] C. M. Bishop, Pattern recognition and machine learning. New York: Springer, 2006.
[21] S. Marsland, Machine learning: an algorithmic perspective. CRC press, 2015.
[22] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny, “An Empirical Comparison of Machine Learning Models for Time Series Forecasting,” Econometric Reviews, vol. 29, no. 5–6, pp. 594–621, Aug. 2010, doi: 10.1080/07474938.2010.481556.
[23] C. C. Aggarwal, Outlier Analysis. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-47578-3.

[24] S. Buffa, M. H. Fouladfar, G. Franchini, I. Lozano Gabarre, and M. Andrés Chicote, “Advanced Control and Fault Detection Strategies for District Heating and Cooling Systems—A Review,” Applied Sciences, vol. 11, no. 1, p. 455, Jan. 2021, doi: 10.3390/app11010455.
[25] Y. Sun, F. Haghighat, and B. C. M. Fung, “A review of the-state-of-the-art in data-driven approaches for building energy prediction,” Energy and Buildings, vol. 221, p. 110022, Aug. 2020, doi: 10.1016/j.enbuild.2020.110022.
[26] J. Peppanen, Xiaochen Zhang, S. Grijalva, and M. J. Reno, “Handling bad or missing smart meter data through advanced data imputation,” in 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Minneapolis, MN, USA, Sep. 2016, pp. 1–5. doi: 10.1109/ISGT.2016.7781213.
[27] I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani, “A review of missing values handling methods on time-series data,” in 2016 International Conference on Information Technology Systems and Innovation (ICITSI), Bandung - Bali, Indonesia, Oct. 2016, pp. 1–6. doi: 10.1109/ICITSI.2016.7858189.
[28] C. Wang, Y. Du, H. Li, F. Wallin, and G. Min, “New methods for clustering district heating users based on consumption patterns,” Applied Energy, vol. 251, p. 113373, Oct. 2019, doi: 10.1016/j.apenergy.2019.113373.
[29] Y. Du, C. Wang, H. Li, J. Song, and B. Li, “Clustering Heat Users Based on Consumption Data,” Energy Procedia, vol. 158, pp. 3196–3201, Feb. 2019, doi: 10.1016/j.egypro.2019.01.1010.
[30] D. Koschwitz, E. Spinnräker, J. Frisch, and C. van Treeck, “Long-term urban heating load predictions based on optimized retrofit orders: A cross-scenario analysis,” Energy and Buildings, vol. 208, p. 109637, Feb. 2020, doi: 10.1016/j.enbuild.2019.109637.
[31] D. Koschwitz, J. Frisch, and C. van Treeck, “Data-driven heating and cooling load predictions for non-residential buildings based on support vector machine regression and NARX Recurrent Neural Network: A comparative study on district scale,” Energy, vol. 165, pp. 134–142, Dec. 2018, doi: 10.1016/j.energy.2018.09.068.
[32] Z. Wei et al., “Prediction of residential district heating load based on machine learning: A case study,” Energy, vol. 231, p. 120950, Sep. 2021, doi: 10.1016/j.energy.2021.120950.
[33] E. Calikus, “A data-driven approach for discovering heat load patterns in district heating,” Applied Energy, p. 15, 2019.
[34] Y. Hong, S. Yoon, Y.-S. Kim, and H. Jang, “System-level virtual sensing method in building energy systems using autoencoder: Under the limited sensors and operational datasets,” Applied Energy, vol. 301, p. 117458, Nov. 2021, doi: 10.1016/j.apenergy.2021.117458.
[35] C. Wang et al., “Research on thermal load prediction of district heating station based on transfer learning,” Energy, vol. 239, p. 122309, Jan. 2022, doi: 10.1016/j.energy.2021.122309.
[36] W. Zhong, E. Feng, X. Lin, and J. Xie, “Research on data-driven operation control of secondary loop of district heating system,” Energy, vol. 239, p. 122061, Jan. 2022, doi: 10.1016/j.energy.2021.122061.
[37] M. Gong, Y. Bai, J. Qin, J. Wang, P. Yang, and S. Wang, “Gradient boosting machine for predicting return temperature of district heating system: A case study for residential buildings in Tianjin,” Journal of Building Engineering, vol. 27, p. 100950, Jan. 2020, doi: 10.1016/j.jobe.2019.100950.
[38] A. Golla, J. Geis, T. Loy, P. Staudt, and C. Weinhardt, “An operational strategy for district heating networks: application of data-driven heat load forecasts,” Energy Inform, vol. 3, no. S1, p. 22, Oct. 2020, doi: 10.1186/s42162-020-00125-5.
[39] B. Ciapała, J. Jurasz, M. Janowski, and B. Kępińska, “Climate factors influencing effective use of geothermal resources in SE Poland: the Lublin trough,” Geotherm Energy, vol. 9, no. 1, p. 3, Dec. 2021, doi: 10.1186/s40517-021-00184-1.
[40] G. Xue, J. Song, X. Kong, Y. Pan, C. Qi, and H. Li, “Prediction of Natural Gas Consumption for City-Level DHS Based on Attention GRU: A Case Study for a Northern Chinese City,” IEEE Access, vol. 7, pp. 130685–130699, 2019, doi: 10.1109/ACCESS.2019.2940210.

[41] C. Wang, J. Yuan, J. Zhang, N. Deng, Z. Zhou, and F. Gao, “Multi-criteria comprehensive study on predictive algorithm of heating energy consumption of district heating station based on timeseries processing,” Energy, vol. 202, p. 117714, Jul. 2020, doi: 10.1016/j.energy.2020.117714.
[42] B. Marlin, “Missing data problems in machine learning.,” Library and Archives Canada = Bibliothèque et Archives Canada, Ottawa, 2010.
[43] A. S. Ahmad et al., “A review on applications of ANN and SVM for building electrical energy consumption forecasting,” Renewable and Sustainable Energy Reviews, vol. 33, pp. 102–109, May 2014, doi: 10.1016/j.rser.2014.01.069.
[44] K. Nam, S. Hwangbo, and C. Yoo, “A deep learning-based forecasting model for renewable energy scenarios to guide sustainable energy policy: A case study of Korea,” Renewable and Sustainable Energy Reviews, vol. 122, p. 109725, Apr. 2020, doi: 10.1016/j.rser.2020.109725.
[45] N. Bokde, M. W. Beck, F. Martínez Álvarez, and K. Kulat, “A novel imputation methodology for time series based on pattern sequence forecasting,” Pattern Recognition Letters, vol. 116, pp. 88–96, Dec. 2018, doi: 10.1016/j.patrec.2018.09.020.
[46] R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui, “A comparative evaluation of outlier detection algorithms: Experiments and analyses,” Pattern Recognition, vol. 74, pp. 406–421, Feb. 2018, doi: 10.1016/j.patcog.2017.09.037.
[47] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation Forest,” in 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, Dec. 2008, pp. 413–422. doi: 10.1109/ICDM.2008.17.
[48] JooSeuk Kim and C. Scott, “Robust kernel density estimation,” in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, Mar. 2008, pp. 3381–3384. doi: 10.1109/ICASSP.2008.4518376.
[49] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, “Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data,” in Advances in Knowledge Discovery and Data Mining, vol. 5476, T. Theeramunkong, B. Kijsirikul, N. Cercone, and T.-B. Ho, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 831–838. doi: 10.1007/978-3-642-01307-2_86.
[50] D. M. Blei and M. I. Jordan, “Variational inference for Dirichlet process mixtures,” Bayesian Anal., vol. 1, no. 1, Mar. 2006, doi: 10.1214/06-BA104.
[51] R. Wu and E. J. Keogh, “Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress,” in 2022 IEEE 38th International Conference on Data Engineering (ICDE), May 2022, pp. 1479–1480. doi: 10.1109/ICDE53745.2022.00116.
[52] S. Sarfraz, V. Sharma, and R. Stiefelhagen, “Efficient Parameter-Free Clustering Using First Neighbor Relations,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 8926–8935. doi: 10.1109/CVPR.2019.00914.
[53] T. R. Bott, Fouling of heat exchangers. Amsterdam ; New York: Elsevier, 1995.
[54] T. L. Bergman, A. Lavine, and F. P. Incropera, Fundamentals of heat and mass transfer, Eighth edition. Hoboken, NJ: John Wiley & Sons, Inc., 2017.
[55] A. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring Network Structure, Dynamics, and Function using NetworkX,” p. 5, 2008.
[56] B. Falay, G. Schweiger, K. O’Donovan, and I. Leusbrock, “Enabling large-scale dynamic simulations and reducing model complexity of district heating and cooling systems by aggregation,” Energy, vol. 209, p. 118410, Oct. 2020, doi: 10.1016/j.energy.2020.118410.
[57] G. Schweiger, P.-O. Larsson, F. Magnusson, P. Lauenburg, and S. Velut, “District heating and cooling systems – Framework for Modelica-based simulation and dynamic optimization,” Energy, vol. 137, pp. 566–578, Oct. 2017, doi: 10.1016/j.energy.2017.05.115.
[58] M. H. Kristjansson and M. B. Bøhm, “Optimum Design of Distribution and Service Pipes,” p. 11.
[59] P. Wallentén, “Steady-state heat loss from insulated pipes.” 1991. [Online]. Available: https://lucris.lub.lu.se/ws/files/4836934/8146384.pdf
[60] “LOGSTOR District Energy,” Product Catalogue, Mar. 2020. [Online]. Available: https://www.logstor.com/media/6507/produktkatalog-de-202003.pdf
[61] M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” Sci Data, vol. 3, no. 1, p. 160018, Dec. 2016, doi: 10.1038/sdata.2016.18.

8. Appendix A

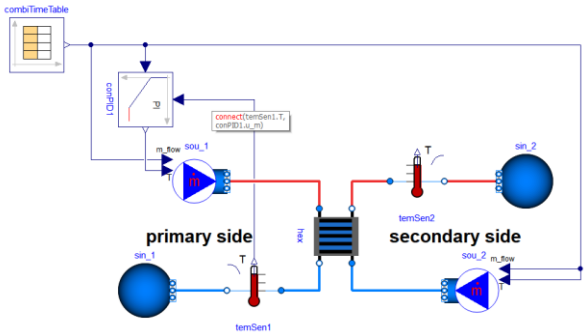
8.1. Software repositories

During the project, the following software repositories were created

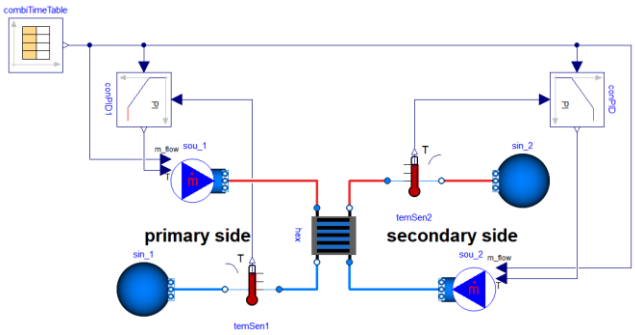
- Data cleaning pipeline: https://doi.org/10.5281/zenodo.7470102
- HT_sim: https://doi.org/10.5281/zenodo.7510252
- DH_sim: Not publishable due to dependence on proprietary models

8.2. Dymola models for HT_sim method

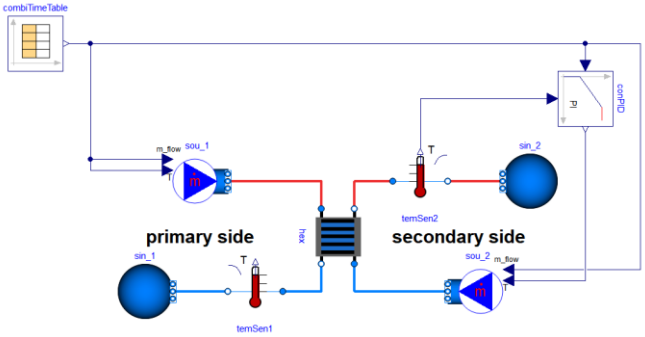
The following Dymola models for missing data cases covered by the HT_sim method (see Table 4) were created, the underlying code can be found in the repository linked in Section 8.1:



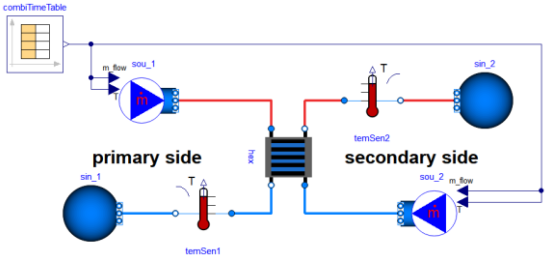
PlateHEXepsNTU_T1in



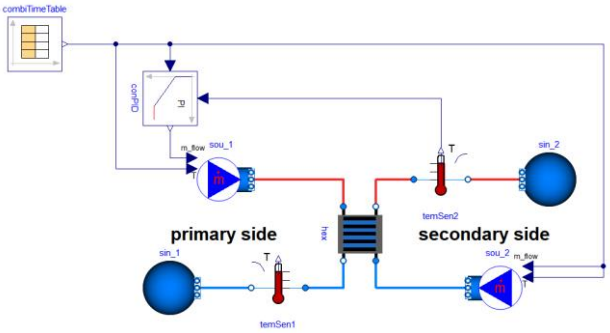
PlateHEXepsNTU_T1inT2in



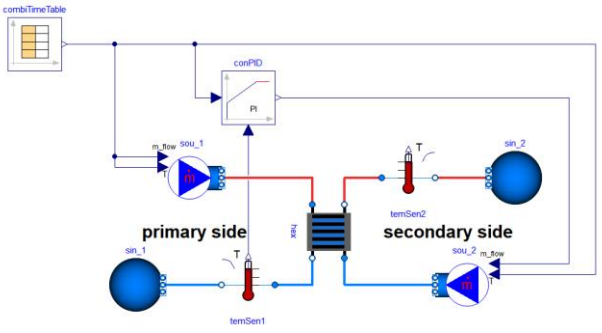
PlateHEXepsNTU_T2in



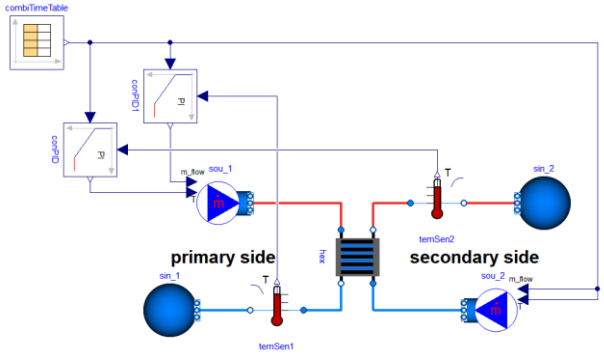
PlateHEXepsNTU_fieldTest



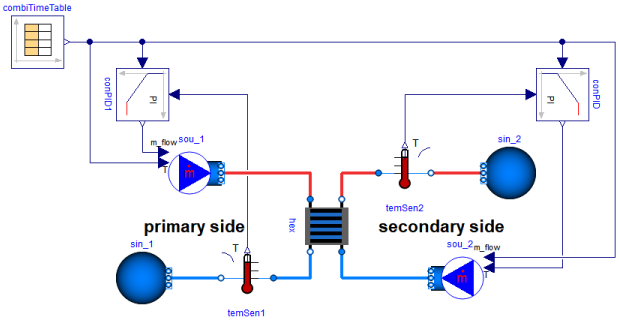
PlateHEXepsNTU_m1



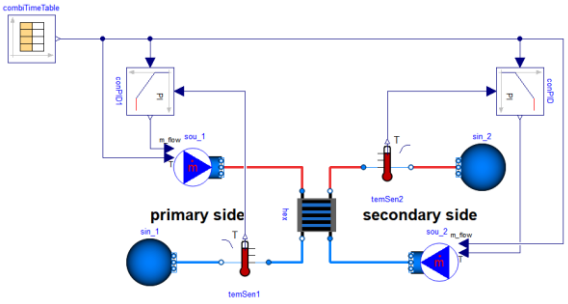
PlateHEXepsNTU_m2



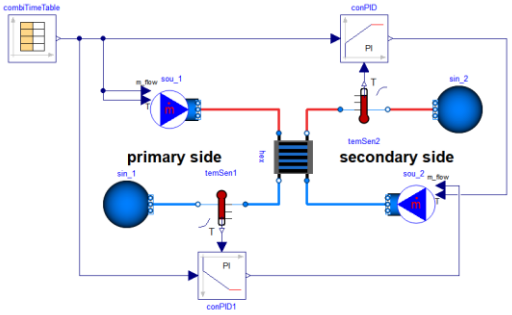
PlateHEXepsNTU_m1T1in



PlateHEXepsNTU_m2T1in



PlateHEXepsNTU_m1T2in



PlateHEXepsNTU_m2T2in

9. Appendix B

9.1. List of Abbreviations

Abbreviation	Description
ANN	Artificial neural network
AR	Autoregressive model
ARIMA	Autoregressive integrated moving average model
ARMA	Autoregressive moving average model
BBN	Bayesian belief network
BPNN	Back propagation neural network
CNN	Convolutional neural network
DH, DHS, DHN	District heating, District heating system, District heating network
DHN_sim	District heating network simulation (physics-based missing data imputation method)
DNN	Deep neural network
DPP	Data pre-processing pipeline (anomaly detection and missing data imputation method)
DPP_HT	Sequential application of DPP and HT_sim (combined method)
DPP_HT_DHN	Sequential application of DPP, HT_sim and DHN_sim (combined method)
DT	Decision Tree
ELM	Extreme learning machine
ES	Exponential smoothing
EWMA	Exponential weighted moving average
FINCH	Efficient parameter-free clustering using first neighbor relations
GRU	Gated-recurrent units
HT_sim	Heat transfer station simulation (physics-based missing data imputation method)
LR	Linear regression
LSTM	Long short-term memory
LWMA	Linear weighted moving average
MA	Moving average model
MAE	Mean absolute error
MAR	Missing at random
MCAR	Missing completely at random
ML	Machine learning
MLR	Multi linear regression
NA	Not available (missing value)
NARX	Non-linear autoregressive with exogenous inputs
NMAR	Not missing at random
RF	Random forest
RT	Regression tree
SMA	Simple moving average
SVM	Support vector machine
SVR	Support vector regression
WNN	Wavelet neural network
XGBoost	Extreme Gradient Boosting

9.2. List of Figures

9.3. List of Tables

10. Contact

Project manager

Marnoch Hamilton-Jones, MEng
m.hamilton-jones@aee.at
Phone: +43 (0) 3112-5886-0, ext. 226

Institute

AEE – Institute for Sustainable Technologies (AEE INTEC)
Feldgasse 19, A-8200 Gleisdorf
Fax: +43 (0) 3112-5886-18
www.aee-intec.at/

Project partner

Know-Center GmbH Research Center for Data-Driven Business & Big Data Analytics