Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

INCI YUEKSEL-ERGUEN, JANINA ZITTEL, YING WANG, FELIX HENNINGS, THORSTEN KOCH

Lessons learned from gas network data preprocessing

ZIB Report 20-13 (July 2024)

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30-84185-0
Telefax: +49 30-84185-125

E-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Lessons learned from gas network data preprocessing

Inci Yueksel-Erguen[1], Janina Zittel[1], Ying Wang[2], Felix Hennings[3], and Thorsten Koch[1,3]

[1] *Zuse Institute Berlin, yueksel-erguen@zib.de; zittel@zib.de; koch@zib.de*
[2] *I²DAMO Intelligente Integrierte Datenanalyse und Mathematische Optimierung GmbH, wang@i2damo.de*
[3] *Chair of Software and Algorithms for Discrete Optimization, Institute of Mathematics, Technische Universität Berlin, hennings@math.tu-berlin.de*

July 17, 2024

Abstract

Due to the critical role of natural gas in the European energy system, its secure and efficient transport is mandatory. Recent disruptive events have drastically changed the pan-European gas transport. To deal with the impacts, algorithmic intelligence methods solving cumbersome real-world problems are crucial. Yet, their results are only as reliable as the input data. However, preparing high-quality data is surprisingly challenging. One can hardly overestimate the necessary effort. In the last decade, we have spent many person-years on the German natural gas transport network data, including data on its technical elements such as compressors. The network consists of thousands of interconnected elements spread over at least 120,000 km of pipelines built during the last century. Our studies revealed the criticality of a suitable gas network data preprocessing (GNDP) strategy enabling adequate data. Here, we present the lessons learned for GNDP from our studies on this network through two case studies. Besides, we describe various GNDP strategies to overcome the obstacles and minimize future endeavors. In the first case study, we demonstrate GNDP strategies when employing data from our industry partner. We emphasize that the well-compiled data sets enhance the company's capability to respond to unprecedented cases. In the second case study, we rely on open data. Here, we highlight high-quality research data sets. They establish the foundations for innovative algorithmic decision-making solutions in gas networks to future challenges.

Keywords— GasLib, gas transport, networks, stationary / transient gas network optimization, real-world data consistency, data preprocessing

1 Introduction

The reliability of the decisions based on algorithmic intelligence methods relies on the underlying data quality. Data is ubiquitous in the age of analytics. However, identifying pertinent data and assessing its quality is demanding. It is incredibly

costly to construct or assemble a high-quality data set. Hence, comprehending data needs is crucial for developing algorithmic intelligence methods that yield reliable decisions. These points are especially true for gas network data representing a highly-interconnected infrastructure belonging to multiple owners and having some legacy parts installed before digitization age.

A complete isolation of data curation and algorithm development in applied studies is unrealistic when working with gas networks. Practically, researchers have to preprocess data before and during the algorithm development. However, one can easily get lost in gas network data preprocessing (GNDP). This task is intricate. The intricate structure of gas network components, such as compressor stations, and multivariate data involving digitized network topology and the technical properties of the components even complicate the GNDP. Besides, there is a plethora of available tools and methods addressing general data preprocessing. Therefore, the adopted GNDP strategy is paramount for the time and cost-efficiency of the algorithm development process as well.

The main contribution of this study is a toolbox for GNDP. This toolbox includes a guide for strategy selection to improve gas data set quality. We also propose utilizing mathematical modeling and optimization know-how in GNDP. We demonstrate our implementation results and lessons learned on the German gas transport network data set. We employ case studies for demonstration. We share our insights especially when data improvement should continue parallel to the modeling and algorithm development. Besides, we show the need for domain expertise and modeling know-how to construct well-compiled open research data sets. These data sets serve as a benchmark for novel algorithmic solutions.

As a result of the study, we published the network visualization software PyNet (Python Network Visualization Tool) publicly [Zus22]. In addition, we managed to compile a research data set for German gas transport network data. This data set is consistent with the high-level European gas network topology data published by the European Network of Transmission System Operators for Gas (ENTSOG) [ENT19, YEMW+23].

In the paper, initially, we summarize the related work in Section 2. Then, we deliver the detailed problem definition in Section 3. We present the data preprocessing methods with a guide for adopting GNDP strategy in Section 4. We demonstrate the proposed guide in two case studies involving gas network optimization. In Section 5, we summarize the data requirements for this context. The case studies address organizational data and open data cases. They are presented in Sections 6 and 7, respectively. Here, we present how we cope with the diverse application challenges by employing various GNDP strategies. We provide some guidance on how to deal with the most common obstacles. We elaborate on the lessons we learned from these case studies in Section 8. In the last section, we discuss the value added by well-compiled data sets in the gas network optimization context and why preparing them is worth immense effort.

2 Related Work

2.1 Public research data sets based on real-life data

Testing novel algorithmic intelligence-based decision-making approaches require data sets. If an adequate benchmark data set is available for testing, its quality is often taken for granted. Therefore, the quality of such data sets influences the conclusions drawn from these studies.

Compiling comprehensive research data sets from real-world data is intricate. Research data sets are not always constructed from scratch. In some cases, data sets

are built using existing data designed for a manifold of purposes [OWPT10]. Sometimes, data reside in separate systems without any connections between them [AGS19]. Some data systems have grown historically without a joint plan from the beginning [BBC+21].

There are multiple open research data sets for gas network optimization. They are derived from various research questions. These examples can illustrate the complications of research data set construction. Based on their varying scope and detail, we note that the complications can be extended to other application domains.

Only a limited number of real-world physical gas network data sets as [DWS00] are publicly available. Efforts to build an open gas network data set focus either on just one of the economic and physical features of the gas network, or on both.

In the former case, data sets may include only network topology. [CBB+09] uses predefined data from a data supplier and focuses on missing parts of the EU network topology. They use methods for substituting incomplete data, like pipeline capacities. [BGH+19] utilizes heuristics and assumptions based on gas network mathematical modeling knowledge. GasLib also provides data sets of real gas networks, but data is distorted due to confidentiality requirements [SAB+17]. Some studies focus only on the demand and supply part of the data. For example, [HHH+15] employs heuristics to generate adversarial nomination data for the gas network optimization models. This data allows stress testing of the network.

The latter case is rather holistic. The resulting data sets can be input to an end-to-end analysis of gas transport networks that use the gas demand and supply to output the gas flow. Among those, [KKS+17, Li12] consider the German gas network infrastructure. [Li12] claims that their methodology can be extended to the European gas transport network. Besides, [Sci18] addresses a data set for the European gas transport network. [Roe14] also includes economic data and network components, e.g., compressor station data for Germany. However, it excludes the complete network topology data.

GasLib data model comprises the required data for a stationary gas network optimization model that allows detailed compressor station modeling [SAB+17]. GasLib data format is based on XML (Extensive Markup Language). Besides, a schema definition in XSD(XML Schema Definition) format is available [Zus18a].

Consequently, the available open gas network data sets are either incomplete compared to the GasLib data model, or inaccurate. Therefore, they must be improved before being employed by stationary or more precise gas network optimization models to make consistent real-world decisions. Hence, data quality improvement and assessment are necessary before

- employing these data sets in gas network optimization using stationary or more precise models, or,
- publishing their updated versions as high-quality benchmark test instances.

2.2 Data quality and data improvement

The paper emphasizes leveraging existing data to compile better-quality data sets, especially in the context of optimizing gas networks, rather than building them from scratch. It stresses the significance of general data preprocessing techniques [HN20] in bridging the gap between available data and the data needed for our models. In addition to data preprocessing, it also highlights the necessity of data quality assessment and improvement techniques to ensure that the compiled data sets meet the required quality for our decision-making models [EW22, AT22, RMS+22].

Data quality assessment has been a focus for many years, and recent advancements in automated data improvement techniques have gained momentum due to the increasing volume of data and the rise of machine learning [WSF95, WS96, SLW97,

BCFM09, EW22, AT22, RMS$^+$22, IN22, BGR$^+$15, CZ15]. Defining data quality in dimensions and measuring them with metrics is a common approach in both research and practical applications [EW22, BS20, ACD$^+$13, HHK$^+$18].

The abundance of data-quality literature and tools underscores its importance. However, there is no one-size-fits-all approach for data assessment and improvement. Therefore, understanding data quality concepts and applying appropriate preprocessing techniques are critical for effective GNDP.

3 Problem Definition

The reliability of decisions yielded by algorithmic intelligence counts on the underlying data quality. Particularly when dealing with real-world problems including details of the gas network operation, it is unavoidable to employ highly-connected and consistent real-world gas network data sets to model complex decisions. However, identifying pertinent accessible data and assessing its quality is challenging. It requires understanding the data quality issues and utilizing the data preprocessing methods as necessary to remove these issues. In this section, we describe why we need to deal with data quality issues in general when working with infrastructure data and challenges in data quality assessment and improvement for highly connected infrastructure networks.

First of all, when working with infrastructure networks such as gas networks, a data set may not readily exist. In some cases, some of the required data may be inaccessible, or data may be scattered around to various data sources. For example, for the former, data of a legacy system may only exist on a high level that once was sufficient. Another example is inaccessible data in the public domain. The commercial value of data or security risks is a potential reason for its confidentiality. For the latter, to have a comprehensive data set, one should collect data from various sources. However, these sources do not always have a uniform format or semantic definition framework. A thorough understanding of the data semantics is necessary to integrate the collected data consistently. To investigate and find a relevant piece of data within the large volume of available data is also an issue [MDR15]. The use of online or unstructured data makes it even more complex. Such problems are valid not only for researchers who require open data. They are also relevant for organizations having multiple systems to collect and store data.

When we have an admissible data set to start the analysis, we must still understand its sufficiency for reliable decision-making. Data preprocessing to improve the data may be needed. The quality dimensions provide a systematic way to approach data quality assessment. In this study, we address six data quality dimensions in line with the definitions of the Data Management Association (DAMA) [ACD$^+$13].

- *Accuracy* is the ability to reflect reality. An accurate data set can describe real-world objects correctly. It is not straightforward to measure how accurate data is. The difficulty lies in identifying a reference reality from which the data deviates. Measurement errors from sensor data may inject inaccuracy. Again, insufficient detail given the modeling requirements, wrong data modeling assumptions, and data integration errors are other reasons for inaccurate data sets.
- *Completeness* is the existence of all required data in the data set. A complete data set is comprehensive. Missing entities and missing attributes are reasons for incompleteness.
- *Consistency* means conflict-free data. In a consistent data set, data attributes invariably address the physical entities. Integrating data from different data sources may cause inconsistency, mainly because of the conflicting data due to integration errors like misinterpreting the source data attributes. Again, errors in the source data sets, like data entry errors, cause inconsistency.

- *Timeliness* shows whether data is up-to-date or not. It is mostly relevant for time series data, like the amount of commodity flow on the network. Yet, it is also critical for infrastructure-related data. For instance, a data set may ignore the planned expansions that are inaccurate or inaccessible [YEMW+23]. Similarly, the available data for some facilities may become invalid over time. Potential reasons are machine wear or maintenance activities.
- *Uniqueness* implies representing each real-world object in the data set exactly once. Repeated use of names or duplicate entries for an entity with different names are examples of non-uniqueness.
- *Validity* is conformance to predetermined type, range, and format. Typos in the names of entities and non-conformance to bounds are some examples of invalidity.

In the literature, measuring the data quality with metrics is a common practice (please see Section 2). However, the metrics used to evaluate the dimensions are highly context-dependent [EW22]. Another related practice is to provide data quality as an attribute in the data set for the data users. For reliable decisions, it is still crucial to understand the meaning of the data quality information in the application context [FCSB03, PS08].

Just like other infrastructure network data, data quality assessment and improvement is not a one-time task for real-world applications involving gas networks. Conversely, it is an incremental task and continues even after the analysis starts. In our studies involving gas network data, which is highly-interconnected, some analysis encountered data errors that are too complex for humans to understand. While detecting such errors is a significant achievement, removing them is extremely difficult. Besides, articulating them with the data owners is challenging.

4 Data Preprocessing Methods

With sufficiently high data quality, one can choose a model that defines the problem correctly. Then, we can start developing models and algorithms with high-quality, real-life data sets. However, in practice, this expectation is not realistic.

When we do not have an adequate data set, the problem definition should derive a correct decision analytics process. Hence, in that case, we should account for the decision requirements first instead of the available data (see Figure 1). Thus, if the available data is of insufficient quality, we should adopt strategies to preprocess the available data. So, we improve its quality instead of degrading the model requirements.

To understand the sufficiency of data quality, we should scrutinize the modeling requirements first. Then, we map the decision requirements to the modeling requirements. We use the following three dimensions to check the sufficiency of data for a model fulfilling these requirements.

- **Modeling detail** increases with the number of system elements in the model. We contemplate elements to retain in the model by assessing their effect on the decision. Selecting an inferior model in that sense cause a failure to encounter relevant answers. In contrast, selecting a more detailed model than needed introduces unnecessary hardship in finding necessary data or solving the model. Moreover, detailed models require mathematical precision. As the detail level increases, the number of components addressed in the model increases, so as the required number of data attributes.
- **Computational size** grows with the model's spatio-temporal span and resolution. Using an adequate size is critical to find non-trivial answers. Even if the data is available, employing large-span and high-resolution is not always

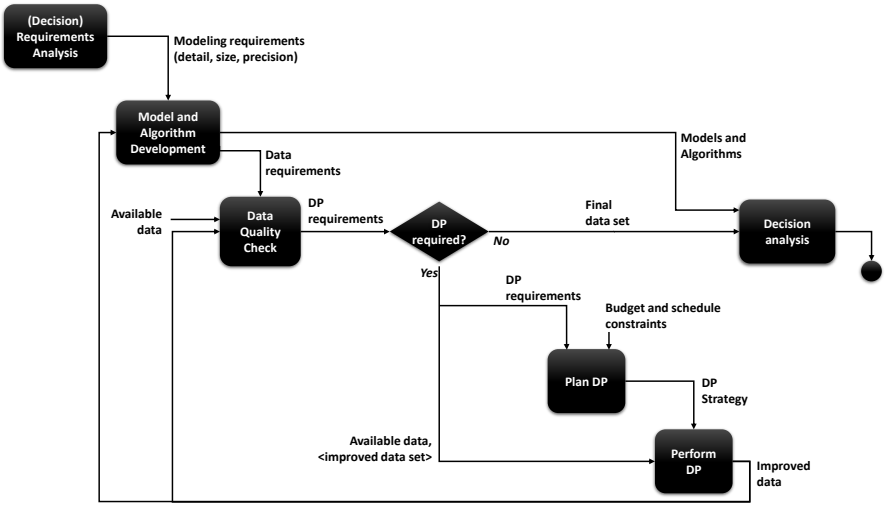


Figure 1: Data Preprocessing (DP) to Deal with Real-World Data

possible. For example, the models may become unsolvable due to the unnecessary large size. Again, a large span may result in solving independent systems together. An adequate model may require aggregation or disaggregation of the available data.

- **Mathematical precision** of the model is its fidelity to the physical properties of the modeled system. For mathematical optimization, mathematical precision is in the increasing order for linear programming (LP), mixed integer linear programming (MILP), mixed integer nonlinear programming (MINLP), and constraint programming (CP).

Consequently, data preprocessing is an intricate task that requires a considerable amount of time. However, after we begin modeling and algorithm development, we often have only limited time and budget for data preprocessing. Hence, adopting a proper strategy at this phase is crucial.

In this study, we present a toolbox for GNDP methods to facilitate strategy selection for gas network analysis. We group the methods into strategy classes according to their level of complexity that induces application difficulty. Although presented for gas network studies in this paper, it is important to note that the toolbox is also applicable to other infrastructure networks given the resemblance of the complexities induced by the required data sets for modeling.

S1. Data cleaning - small interventions using generic tools and methods: The most common approach in mathematics is to employ the data as given with small interventions. This strategy is common when reusing an existing data set. Once a data set has been employed in research, subsequent users rarely question its quality. However, independent from the past uses of a data set, we propose using this strategy when:

- the available data is structured [BS20] such as tables in a relational database or a set of instances defined according to a schema definition, and
- the available data is sufficient to run the model with the required modeling detail, computation size, and mathematical precision.

Then, context-agnostic data-cleaning methods can be sufficient. Some examples of these methods include outlier filtering [KKP07, OGP06], deleting redundant data, and replacing missing values by substitute value formation [CBB+09, KKP07, KDS12]. These methods neither require context-dependent settings nor additional information from any external resource. In other words, we do not need additional data from other sources or knowledge from sources like subject-matter-experts.

S1 addresses data accuracy and completeness improvement by schema validation, outlier detection, or data imputation. Similarly, investigating inconsistencies in data by methods like data integrity check is an S1-type method. There are also automated tools for such manipulations as listed in some recent survey papers [EW22, HN20, AT22]. However, most of these tools are not open-source. They are mainly commercial.

If the data set is multivariate, i.e., includes both spatial and tabular data, the geometric or topological relations of the data attributes also matter. Employing tools specific to spatial data is required for data preprocessing. These tools can check spatial data errors based on generic geometric relationships. Again, they may have a rule-based error-checking methodology [FAO+20]. In this case, the data owner uses predetermined rules. However, it is not always possible to access the digitized network data or the data preprocessing tools when they are commercial. Hence, heuristics that exploit the generic geographical or topological properties of spatial data such as connectedness or triangular inequality [Li12] are also methods in S1.

S2. Data modeling - for (further) standardizing input data: If the input data is subject to change over time, using data preprocessing, we aim to anticipate problems caused by data updates. In this case, data preprocessing includes procedures to deal with data inconsistencies before they even arrive [RKCMP19]. This strategy is essential for systems that deal with online data. Similarly, they are important for the reusability and maintainability of the benchmark research data sets.

An example of these methods is to define a schema including the naming, type, and range specifications of the data [AGN15] as in the case of GasLib [Zus18a, SAB+17]. The schema definitions help maintain the uniqueness and validity of data. Data standardization is also relevant for tasks involving data governance frameworks in organizations [AGN15, JBE+20]. S2 also incorporates updating existing data relations or schemas when dealing with structured data. Similarly, it is employed to define schemas of semi-structured data such as XML records.

S3. Data enrichment - using resources external to the data set: We adopt this strategy when the available data is insufficient to meet the modeling needs since the available data does not fit

- *modeling detail*: This problem occurs when data for some system components retained by the model partially exist or do not exist. This case differs from the case of the missing values in S1 since structured data is addressed by S1. In that case, data imputation methods may be sufficient to fill in the missing data, i.e., by exploiting data integrity rules. Another reason is, the existing data may be insufficient granularity or precision to match the mathematical precision needed by detailed modeling. Such a data set is not accurate and complete enough to be used in decision-making.
- *the geographical span/resolution, or the temporal span/resolution*: In this case, data should be either aggregated or disaggregated. In both situations,

various approaches as mathematical models and statistical methods, are convenient. In some cases, external data, i.e., to serve as meta-distribution, can be employed [YEMW$^+$23].

In such cases, either data is not completely available, or the effort of compiling it is beyond the project scope. To deal with such problems, we enrich data by

- *data augmentation*, i.e., by enriching available data set using adequate data from external data sources [AT22, WOB14, HN20]. In this context, by augmenting data, we mean adding some data attributes to the existing data set or completing its missing values using another data set. One must first collect data from suitable sources. Then the collected data is integrated into the existing data set. During integration, relevant logical or mathematical relationships are considered as necessary. Here, comprehension of the semantics and purpose of the other data source is of utmost importance to avoid data inconsistencies.
- *data generation*, i.e., by making educated assumptions, using heuristics, or mathematical models [CBB$^+$09, HHH$^+$15, BGH$^+$19, BBC$^+$21, Li12]. Data generation is often required to employ high-precision models while the available data is incomplete and inaccurate [HSW21, HW18, RKD$^+$13, QSP$^+$22, BGK$^+$14, VIV$^+$12, YEMW$^+$23]. Spatial and temporal data aggregation/disaggregation using mathematical models to fit data for the required resolution is an example of these methods.

Such methods are context dependent. They require mathematical modeling knowledge as well as application domain knowledge. Besides, tailoring is necessary for individual research and development purposes to ensure reliable end products. Moreover, these methods may inject inconsistencies into the data set. If this is the case, detection of them by humans or simple consistency check heuristics may not be possible. Thus, the methods in S3 require more resources and time compared to other strategies.

S4. Data error diagnosis and correction - by exploiting knowledge external to the data set: This strategy addresses the cases where data is admissible enough to start the decision analysis but still not completely correct. Such data sets have errors missed by the applied data preprocessing methods or even injected by them. Errors reveal themselves in the results of the decision analysis, e.g., with systematically infeasible or inconsistent/unexpected results. Diagnosing and correcting these errors requires a thorough knowledge of how the modeled system operates. One option is to report such errors when detected during research and correct them at the source system level.

The data producer should be involved in the project to employ this strategy. The data producer's involvement is the only acceptable approach to yield reliable results for certain types of very complex data problems. Imagine, for example, a model with data that should be feasible, but an NP-hard computation is necessary to prove this. The amount of data corrected in such an iterative approach is limited. Furthermore, it is typically beyond the scope and budget of the research and development project.

On the other hand, exploiting knowledge from mathematical models may be necessary when working with open data. In such cases, it is not possible to involve the data producer. Thus, evaluating data using a mathematical modeling set-up can provide some insights to diagnose and, in some cases, correct the errors. For highly-connected and highly non-linear data, it is a chief achievement even to detect such errors. Yet, methods like mathematical model preresolving, slack formulations, and minimum irreducible infeasible subsystems can be employed

to further analyze infeasible cases for an understanding of such highly-complex data errors [JSSW15].

The above-defined strategies require different levels of application domain and mathematical modeling knowledge. Besides, they have varying application difficulty. For example, among them, S1 requires the least domain knowledge and mathematical modeling knowledge. Again, to apply S3 or S4, particularly the version that exploits knowledge from mathematical models, require more time and effort compared to others due to their application difficulty. So, S2 or S1 should be preferred whenever they are sufficient. Hence, we propose adopting an incremental implementation of the flow in Figure 1 for decision analysis integrated with GNDP. In this flow, we apply the strategy that meets modeling requirements with the least effort in each increment.

We also propose using visualization tools that allow simultaneously visualizing the data and the modeling results with the use of the data. They are especially effective when we check the data quality against modeling requirements. When data include network structures, graph visualization tools are invaluable in detecting topological errors and unexpected bottlenecks in the modeling results. On the other hand, visualization helps communicate errors to the subject matter experts, especially when using S4 to diagnose and correct errors with the help of the data owner's knowledge.

5 Gas Transport Network Data

Gas networks consist of gas pipelines in which the gas flows. In this paper, we focus on gas transport networks comprised of higher-diameter and higher-pressure pipelines. Their function is to transport gas long-distance, i.e., within a country or continent. Hence, we refer to *gas transport networks* shortly by *gas networks* here after in the paper.

The European energy system is going through a fundamental transition. The aim is to meet greenhouse gas emission targets. In this respect, the pan-European gas network is essential. For instance, the total energy amount transported through German gas pipelines is double the amount transported through the entire electricity grid in Germany [BMW23]. Hence, a consistent evaluation of gas network capacity is crucial for reliable prospective decisions. On the other hand, because of the liberalization of the European gas market in 2009 [Eur09], TSOs do not influence where gas is supplied or stored. Hence, efficient operation of the gas network while maximizing TSOs' capacities to offer is a requirement.

The significance of gas networks has been ever-increasing with requirements induced by technologies like power-to-gas (P2G). Moreover, recent political issues in Europe revealed the criticality of gas to ensure the security of the energy supply. Consequently, the properties of gas and its main flow direction have changed. To understand the impacts of such changes, decision-makers, ranging from policymakers to system operators, need automated decision support systems. These systems employ more complex models to compensate for the gap between the novel situation and the experience from the past. Such systems crucially depend on reliable, coherent, and consolidated data.

5.1 Decision making problems

We can address diverse decision-making problem types related to the gas network within the European gas market. (see Table 2 for example decisions).

- Strategic decision-making problems address policies involving long-term investment and network expansion decisions [YEMW$^+$23, Loc21]. These are relevant for the entire continent or selected countries with a temporal span of multiple years or decades.

- Tactical decision-making problems address the feasibility of the gas transport infrastructure to meet the strategic level requirements. The geographical span of such models is usually limited to a country. Again, their temporal span is typically limited to years with a resolution of hours to days.
- Operational decision-making problems address short-term decisions for decision support to the TSOs. The focus is on appropriate ways to control the commodity flow in the infrastructure by changing the system settings. To exemplify them, we can count compressor station configurations and valve states. The time resolution for short-term planning problems varies from minutes to hours.

5.2 Gas network optimization models

Gas transport is facilitated by increasing or decreasing the pressure at designated points in the network. These pressure differences induce flow, i.e., gas travels from high-pressure to low-pressure areas. Thus, the maximal amount of gas flow over a particular pipeline depends on the size of the maximal pressure differences between its end nodes. We refer to this amount as *capacity*. Capacity depends on a multitude of different factors. Pressure bounds, velocity limits of the pipeline, and states of the nearby network elements are examples of these factors.

The network elements can be passive as pipelines or active as compressor stations, valves, and control valves. The latter can regulate the gas flow in the network with their states. For instance, a closed valve decouples the network part, and a functioning compressor station or control valve changes pressure. So, the pipeline capacity also depends on the active elements.

We can determine the capacity of a pipeline using mathematical models with an adequate data set. Such data should include a detailed representation of the network topology and a description of all the technical properties and logical dependencies of all the involved elements. For the details of the models, please see the work of [KHPS15] on physical gas network models.

Physical gas network models consider the thermodynamic behavior of the gas in pipelines. The steady-state gas network models are stationary, while transient models are time-dependent. The pressure change due to the active network elements are modeled in the physical network models. However, these models do not always consider their combinatorial operational modes.

For operational-level decisions, models considering the time-dependent thermodynamic behavior of gas and the combinatorial nature of the active elements are needed. So, transient gas network optimization models are typically employed [HBHL$^+$21a].

With the use of technologies like P2G, integration of the gas infrastructure with the energy system is in effect. In this regard, decisions related to re-purposing the gas network to route various gas mixtures, including hydrogen or ammonia, are evaluated in such problems. Hence, physical gas network models have been employed for tactical decisions recently. It is typical to model the non-linear pressure drop of gas in pipelines. However, the compressor stations and other regulating devices are either not modeled or modeled in varying detail. Transient models are computationally intractable to work with real-life networks. Hence, these studies commonly employ steady-state gas network optimization models [SLS$^+$21].

Not all decisions require detailed modeling as physical gas network models. For example, a high-level gas network representation may suffice for strategic-level decisions. They model the cumulative gas flow between countries or gas systems. Similarly, the linear (or linearized) gas network models are typical for tactical-level decisions. These models do not consider the thermodynamic behavior of gas and the effects of active elements in gas flow. Instead, they use pre-calculated pipeline capacities.

5.3 Data Requirements

Gas network data requirements alter with the modeling requirements. For example, a high-level gas network model requires only cumulative capacity values between important gas systems. Yet, a stationary gas network model requires physical attributes for each individual elements as well as physical properties of gas flowing through pipelines. To demonstrate the difference, we examine compressor stations based on models provided by [HW18]. We refer to [HW18] for detailed background information and explicit formulas of all mentioned relationships.

Compressor stations increase the pressure in the gas flow direction. They require at least one compressor unit to function. Each unit conveys a single compressor machine and a drive. Here, drives provide the power for compressors. Compressor machines differ in technical compression elements. Compressor units are arranged in either series or parallel. The former allows a higher compression ratio, whereas the latter provides higher throughput in gas flow. Each compressor station has a predefined set of technically possible arrangements called the set of configurations.

Each single compressor machine has a feasible range determined by its operating constraints. This range depends on the compressor type. Each point in this range is associated with an efficiency value influencing the power needed to compress gas at the corresponding conditions. The feasible range is defined by parameters either given as the compressor's technical properties or fitted based on measured values.

Drives, the other component of compressor units, are also of multiple types. Here, the relevant quantities for modeling are the energy consumed to provide a corresponding amount of power and the maximum allowable power. The latter depends on the compressor's current speed and the ambient air temperature.

To summarize, to model the operation of a compressor station, we require the corresponding compressor units, their set of configurations, a characterization of the feasible operating range of each unit, as well as the energy consumption and maximum power functions for the corresponding drives. For tactical-level decision-making, less detailed physical gas network models not including the set of configurations can be employed [LHGM22]. Again, linear and high-level gas network models do not incorporate active elements.

Modeling requirements alter the suitable data improvement strategy, too. For instance, unexpected model behavior can identify data inconsistencies and errors for high-level or linear gas network optimization models. Thus, solving such models parameterized by the data can help improve the data. On the contrary, for physical gas network models [KHPS15, HBHL$^+$21a], the source of encountered infeasibility might involve a large number of network elements and, therefore, parameters. Thus, finding the one causal data error from this vast amount is a real challenge. It is because of the dependence of pressure on the complex and nonlinear interactions of the other network elements. We designate such data sets as highly interconnected and complex.

6 Case Study 1: Company data from Gas Transmission System Operators

In case study 1, we deliver our experience on the specific challenges of a project for which the data producer support is accessible. Here, we build a robust online system [Zus18b]. The system consists of a forecasting unit [PCG$^+$22] and an optimization unit. The former predicts the hourly supply and demand. The latter determines the sufficiently accurate control schedule for the individual network elements [HBHL$^+$21b, HAHB$^+$21].

6.1 Modeling requirements

We work in close cooperation with an industry partner. Their IT systems are the source of gas network data. Here, we aim to provide a real-time algorithmic intelligence-based decision support system for network operators to enable successful gas network operation. The system's output is a set of required network controls involving (i) the states of the active elements given the present network state and (ii) the expected gas flow in and out from the network.

We require current state of network elements to recommend future control measures for each. The state of a network element involves its pressure, flow, and, if it is active, its current operation mode. We also need already known future changes and limitations in the network, which may, for example, be a result of maintenance work. The resulting consequences for the network elements can range from tighter feasible operating ranges to a prescribed mode of operation for active ones. Hence, we can summarize the modeling requirements as follows.

- modeling detail: Single active and passive network elements modeled with high fidelity to their operational constraints. There are >1000 entry and exit points, >6000 pipes, >3000 valves, and >100 compressor units.
- computation size: The spatial span is consistent with the geographical span of our industrial partner's network. It is discretized by the single elements included in the model. The model has an hourly granularity.
- mathematical precision: High enough to capture the nonlinearity induced by the thermodynamic gas properties and the combinatorial nature of the operational states of the active components.

Given these requirements, we employ a transient gas network optimization model. A forecasting system feeds the future supply and demand data to this model. The forecasting system [PCG+22] combines optimization and machine learning to predict future supply and demand at the entries and exits of the network based on historical data and the current state of the network. It is out of this paper's scope.

6.2 Data quality check

The required data resides in several databases and systems of our industry partner. As we started the project, a comprehensive data set of necessary data in a structured manner was not readily available. Moreover, by then, a data model meeting the modeling requirements of a transient gas network optimization model did not exist.

The available data sources have a different network representation, object identification system, and time granularity. Thus, even if the complete IT system is comprehensive, mapping objects between systems is needed. During this mapping and integration of data, injecting inconsistency into the resulting data set is possible.

For instance, the gas network representation of the simulation system is meticulous. However, it lacks knowledge about the dependencies of the modes and configurations of the single active elements. Hence, it is incomplete regarding modeling requirements. On the contrary, an aggregated version of the network represents the network operators' system containing the dependency data. An aggregated network representation has fewer elements on a higher abstraction level. A complex rule-based scripting system, which is a component of the network operator's system, implicitly has the dependency information. This scripting system transforms the operators' commands into modes and configurations on the level of single network elements. So, this knowledge has to be extracted and integrated into the network representation from the simulation system.

On the other hand, there are differences in the time granularity of the data source systems. It leads to a time lag between different data sources resulting in timeliness

and consistency issues. For instance, the mode and configuration values for the active elements taken from one source and their pressure and flow values from another have a time lag of three minutes. The former denotes the initial state. Thus, if an element switches mode, it results in a conflict in its initial state that lasts up to three minutes. It is not possible to change source systems to fix this issue.

Another incomplete data issue is due to the compressor machines' drives. The drives function by gas drawn directly from the network. The energy consumption of these drives is required data for operational-level drive models. However, they are not measured in the system and hence, are unknown.

6.3 Adopted data preprocessing strategy

From the quality check, we see that the available data has to be improved in all six quality dimensions to meet modeling requirements. To have a complete data set, automatically or manually integrating the data from different systems may be employed. Yet, in our case, asking for data from the data owners was more convenient. Thus, we first updated the existing stationary gas network data model [Zus18a, SAB$^+$17] as in S2. Such a data model leads to a common ground for integrating the data from different data sources of the company. We can then employ a data governance-based methodology. However, there were still unknown data for the compressor station drives. Thus, next, we completed the data by data generation of strategy S3.

After completing the data, we fixed the data timeliness issue with a problem-specific heuristic method. Hence, we again adopted the S3 strategy. We obtained an admissibly accurate and complete data set with the GNDP up to this point. However, we need an automated check to test data quality during our decision support. Hence, we use a data-checker to test data quality by rule-based heuristics. Besides, we utilize methods to evaluate the errors during our analysis. We examine the errors with data owners whenever we encounter a data error by adopting the S4 strategy.

6.4 Data preprocessing and lessons learned

We began by updating the existing data model based on GasLib [SAB$^+$17, Zus18b]. We aimed to comply with the data requirements of the transient gas network optimization model. Thus, we examined necessary and potentially beneficial data attributes by their availability and accessibility. We did not aim to have a one-time task for the updating process. Instead, we used an incremental one to deal with the potential uncertainties of the research and development tasks. Therefore, we started with the narrowest data model meeting our modeling requirements. Once our novel modeling approaches required further data attributes, we updated the data model. We came up with a specific schema for data validation. Consequently, we can detect data errors in a standardized, automated fashion.

We did not use an automated process for data collection and integration. Instead, we asked data owners to provide us with data. We employed the schema definition to collect and integrate data during data enrichment. We developed an XML interface resulting from the schema definition effort to facilitate integration. Unambiguous communication between all the parties that provide data is the key to accomplishing such a task with only some tweaks in the schema definition. We communicated with plenty of employees of our industry partner during this task. The employees were from varying backgrounds, ranging from specialists on the source system and transformation requirements, over researchers, to subject matter experts on gas dispatching. Because of the incremental approach that we used, we had several updates to the XML interface. To illustrate the development of our interface definition, we present a general overview of the version history in Table 1.

Table 1: Version history of the XML interface in the project, featuring the number of sub-versions, the number of changes made in each update and the number of corresponding file type used

Interfaces	0.3.*	0.4.*	0.5.*	1.0.*	1.1.*	2.0.*	2.1.*	2.2.*	2.3.*	2.4.*	2.5.*	2.6.*	3.0.*
Sub-versions	1	2	3	1	1	2	1	2	1	1	2	2	3
Changes	-	31	41	10	6	58	17	5	2	3	5	1	20
# file types	13	17	22	22	22	22	24	25	25	25	24	24	21

Thanks to the incremental approach we adopted, we could allow a partial deviation from the strict XML format for those file types, which were currently in development, manually created, or likely to change often for other reasons. To facilitate this step, we employed the YAML format. This format uses minimal syntax and is more human-readable than XML. We also integrated a free-text area into the XML. So, we could store data that is only relevant to ourselves, e.g., to store debugging output or information from previous executions.

In case of a necessary change of the schema definition, we employed automated processes for the adjustment and release management. We found that automation considerably reduces the required effort and decreases the likelihood of errors in those tasks. Furthermore, we created tools to convert data from outdated interface formats to newer ones. Consequently, the data is easily accessible now, and the code for parsing does not need to cover all the previous interface variants.

Data errors are likely to happen, either by errors in the source systems, miscommunication, or bugs in the data transformation code. These errors might lead to failure or unexpected behavior in the subsequent parts of our decision analysis. The causes for these problems in decision analysis are often hard to detect or relate to the original data error. So, we developed a *data checker* to investigate such errors after we validate data against the XML schema.

The data checker automatically checks and corrects errors injected into the data during GNDP. It is a rule-based tool for detecting potential data errors in a systematic and reproducible way. With the data checker, we investigate data errors belonging to more than 100 predefined error types associated with non-conformance to data quality dimensions listed in Section 3. Examples include logical or topological data error types. Inconsistent properties of network elements provided by different data sources cause logical errors. For instance, a pipe with a smaller length than the height difference of its end nodes is not logically possible. Unconnected network elements and loops around a single node are examples of topology errors.

In Figure 2, we present the number of warnings and errors found by the data checker in the regularly delivered input data over multiple months. As seen from the figure, there is a decrease in the number of different error types by roughly two-thirds. It is a result of an interface update to make it more uniform. However, thanks to the update, the data checker became more effective. So, the absolute number of data errors increased significantly.

For each defined error in the data checker, we either have an automated way to handle the error appropriately or abort the analysis. In the former case, we use a GNDP approach of S3 type strategy. In the latter case, we provide an error message describing the source data inconsistency. By communicating the error to our industry partner, as in S4, they were able to find the problems' causes most of the time.

Despite our effort on GNDP, we could not fix all the data quality issues addressed in Section 6.2. In these cases, we focused on heuristic approaches. One example is the initial state inconsistency issue. We adopted an S3-type strategy to solve it. So, we use problematic initial state values as if they are correct and enforce model constraints only for future recommendations. This approach leads to a high number of recommendations for the first time steps if switching to a consistent network state

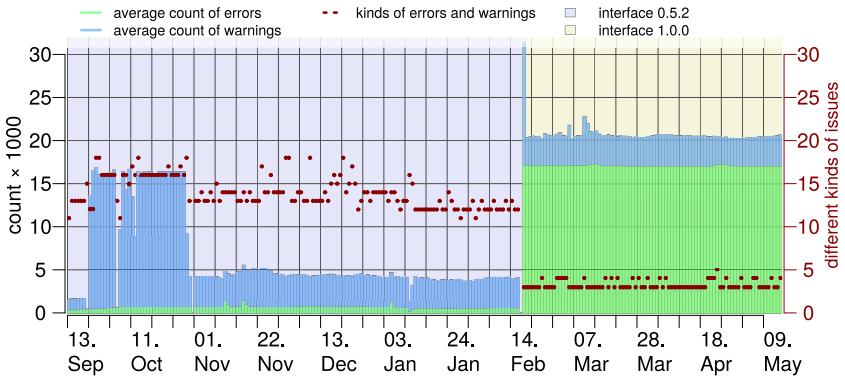


Figure 2: Data errors and warnings found by the data checker in the regularly delivered input data over time. The bars represent the total number of warnings and errors, while the red dots indicate the number of different types of issues(warnings and errors combined). The used interface version can be deduced from the background color

is necessary. However, such a situation reflects the technical problems of the decision-making tool. They do not indicate needed control changes when compared to the dispatchers' control decisions for the same network situation. Hence, the user of the decision-making tool should carefully monitor its results in this regard.

After performing GNDP as we explain above, we achieved an admissibly correct data set. In other words, the data set became accurate and complete and passed all the consistency checks regarding data errors we could have foreseen. However, there were cases where our analysis encountered data errors.

For instance, in one case, we realized that the single compressor machines regularly operated outside their feasible operating range. Figure 3 illustrates an example. According to the subject matter experts, the method employed to define a feasible operation range may lead to such problems. As described in Section 5.3 above, an operation range for a compressor machine is defined based on different measurements. Therefore, this range accounts for neither the operation points attained while starting the compressor machine nor the changes in the machine's properties due to wear. However, it is impossible to regularly repeat the measurements of each compressor machine in the network. Thus, gas network experts of our project partner finally created hand-tailored feasible operating ranges for each compressor machine. Similarly, we got estimates of the energy consumption of gas-powered drives. These kinds of actions are examples of strategy S4 with human interaction.

Using the measures we described here, we were able to employ the data successfully in our decision support system as well as for research projects.

7 Case Study 2: Public data on Gas Networks

In case study 2, we focus on the multi-energy system of Europe in the upcoming decades [CBvAO+21, pla18]. The aim is to understand the restrictions and flexibilities that gas networks induce on the European energy system [MGYE+20]. Case study 2 provides insights from a project relying only on open data.

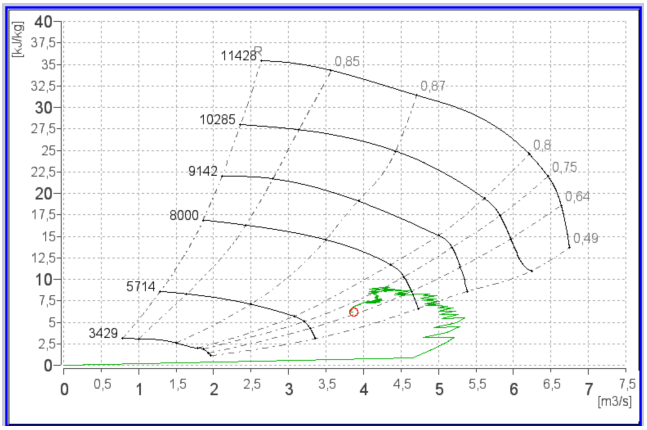


Figure 3: Example of erroneous input data. The grid in black displays the feasible operating range of a compressor unit. The red circle represents the current point of operation while the green line consists of all previously attained operation points. The majority of these previous operation points have been outside of the feasible operating range.

7.1 Modeling requirements

The European energy system has been going through a fundamental transition towards being fossil-free. We explore the operational feasibility of the existing gas network to route future gas flow scenarios. These scenarios represent the gas flow required by the optimal future energy transition pathways to decarbonization.

The significance of gas transport has increased like never before due to the recent political turmoil in Europe. The consequences led to a tremendous impact on the security of the energy supply. Moreover, emerging technologies such as P2G has enabled gas injection other than natural gas, e.g., hydrogen, into the gas network [EEE18, BF18, SGR$^+$15]. Therefore, additional questions arise, such as repurposing some parts of the gas network as hydrogen network [The21] emerge.

Such disruptive events change conventional scenarios, leading to novel situations. So, some experiences from the past become obsolete. Hence, more complex decision models are required to assess the impacts of these events during the energy transition.

Our focus in this study is to investigate the operational feasibility of the German gas network. For analysis, we integrate the gas network optimization model into an electricity grid model with a high spatio-temporal resolution. Therefore, its output, like the electricity grid, should be evaluated hourly at the postal code region level. Our engagement in this project is twofold. First, we prepare high-resolution gas supply and demand scenarios for Germany. Here, we merge high-resolution results from the electricity grid and other gas supply and demand sources. Then, we inspect whether the network has a feasible operational setting to route the resulting scenarios. Thus, the problem definition leads to the following modeling requirements:

- modeling detail: Gas flow in the pipes is slow, unlike electricity flow on wires. Thus, we have to model the physical gas flow for an integrated analysis of gas and electricity grids. Such models comprise single network elements, including the active ones and their operational modes. Besides, we should incorporate the gas supply and demand sources other than the electricity grid in the model. The examples are gas imports and exports of Germany and gas exchange with storage facilities. To properly contain these, we must also consider the pan-European

gas transport constraints.

- computation size: We discretize the geographical span according to the span of the German physical gas network. The time span is the same as an energy transition pathway, e.g., years to decades. However, analyzing days or hours with boundary conditions for a representative year is sufficient for a feasibility study.
- mathematical precision: The model should be precise enough to capture the nonlinearity induced by the gas thermodynamic properties. Again, it should retain the combinatorial nature of the operational states of the complex facilities.

We selected the steady-state gas network optimization model[KHPS15, FGG$^+$15, PGH$^+$15] given the modeling requirements.

7.2 Data quality check

In this case study, we depend on open data. GasLib [SAB$^+$17, Zus18a] defines the required data model for the steady-state gas network optimization model. Hence, in the first step, we searched for open data described in the GasLib data model.

Despite the number of projects working on the European energy transition problem in recent years, there is a shortage of open gas network topology data sets [Ope22a]. The available data sets do not meet GasLib data model requirements (see Section 2.1). Open gas network data at organizations like ENTSOG or national counterparts, for example, the *Vereinigung der Fernleitungsnetzbetreiber Gas e.V.* (FNB) in Germany, only allows high-level economic modeling of the network. TSOs are obliged to publicly publish some information on their gas networks by the European Union [Eur09] and its national implementation of the transparency of gas networks. However, the published information is superficial.

On the other hand, open data sources that we can use to construct our data set are scarce. ENTSOG [ENT18c] provides regular information on gas supply and demands. However, gas network topology data is only available on non-digitized map illustrations [ENT18a]. A data set for the German gas network was published by the research project 'LKD-EU' [LE18, KKS$^+$17]. More recently, SciGridGas [Sci18] data set on a pan-European scale was published. These data sets include network topology with geographical location data, but they are incomplete regarding the GasLib data model.

Data collection and integration from various sources are needed to construct a data set in the detail of the GasLib data model. However, these sources use different standards, naming conventions, and data formats. Besides, they hardly include structured data. For semi-structured data .csv files is a prevalent format, although web data is more common. Moreover, topology data is embedded in images or non-digitized maps. For the few data sources that provide structured data, automatic data collection is mostly banned. There are only a few APIs to download data. Thus, even collecting and formatting the data is a challenge, yet, the data sources are not semantically consistent. For example, points in the TSO data address the entry and the entry nodes, which is intuitive in the gas network optimization context, while points in ENTSOG transparency platform (TP) [ENT18b] are the links connecting different types of gas systems or facilities.

Another issue is the inadequate assignment of geographic location data to the missing network entities, such as compressor stations. TSO and organizational databases do not contain geographic location data, even for pipelines and nodes. They only provide some illustrations that are not georeferenced. Furthermore, pipeline diameters are only available from those visuals. We can make some inferences on pipeline diameters based on structural data on TSO websites, such as cumulative length information of

pipelines per diameter class. On the other hand, Open Street Map (OSM) [Ope17] is far from covering the entire grid, especially for Germany.

For active components, especially for compressor stations, the available data is also insufficient, e.g., compared to our compressor data example in Section 5.3. For instance, the available data for those in Germany is not more than their relative locations to the gas pipes, the number of compressor machines, and maximum power and drive type [Roe14]. Again, this data is incomplete regarding the GasLib data model.

The next challenge is the gas supply and demand forecast data. These are available in ENTSOG databases only as cumulative figures for balancing zones or countries. However, we need the amount of gas entering and leaving the German gas network via entry and exit nodes. As another caveat, the supply and demand forecast has a yearly resolution, whereas we require a temporal resolution of a day or an hour.

Finally, the open data sources only contain pressure bounds for the main pipelines and important nodes of a few TSOs. Therefore, the remaining pressure bounds must be consistently estimated using available data for a more precise analysis.

7.3 Adopted data preprocessing strategy

We started with the data set provided by [KWH$^+$17] instead of constructing a data set from scratch. We aim to improve the existing data set to meet the data requirements of the GasLib data model. First, we augmented this data set with the TSO [FG18, Ope22c, Flu22, GRT22, ONT22, Ter22, Thy22, GTGnd, Gas22] and ENTSOG data [ENT18b] with the help of S1-type tools and models. We also employed some consistency checks with S3-type heuristics. Then, we generated missing active elements' data using partially available data and mathematical models. With the data enrichment, we managed to have an admissible data set for the steady-state gas network optimization model.

Initially, we used S1-type methods for temporal disaggregation of the European supply forecast data, but such methods didn't perform well for spatial disaggregation of the supply data. Instead, we employed S3-type tools to generate scenario data employing mathematical models. With the use of these methods, the quality of scenario data became dependent on the quality of the network topology data.

After we started working with the data using the steady-state optimization model, our analysis encountered errors. So, we exploited the knowledge from mathematical modeling methods to detect and remove such errors. We also used visualization tools to visually analyze the network data as well as the bottlenecks in the result of the analysis using PyNet at any step of the GNDP as necessary.

7.4 Data preprocessing and lessons learned

We used the network topology data provided by project LKD-EU as a basis for our efforts to build a German gas network data set [KWH$^+$17, KKS$^+$17]. First, we determined reliable open data sources for closing the gap between this data set and the GasLib data model. Then, we carefully examined their data-related documentation. In such a task, it is essential to understand how and why the data is collected and published by a data owner. So, we also reviewed related analysis reports and publications. To draw correct conclusions from this study, we had to extensively utilize our know-how in gas markets and gas networks operation and expertise in gas network optimization models. This study laid the foundations of our S3 and S4 type GNDP.

We enriched the base data set to meet the requirements in the GasLib data model. Firstly, we augmented the data using the data obtained from resources external to the data set. Some examples are, data of critical nodes from the TSO and FNB data repositories on transparency web pages [FG18, Ope22c, Flu22, GRT22, ONT22,

Ter22, Thy22, GTGnd, Gas22], European high-level gas network data from ENTSOG [ENT18b], storage facilities data from GIE [Gas21b, Gas21a] and node height data from digitized European Map.

We utilized semi-automated methods of type S1 to collect and integrate data from those resources. We joined node-based data from different data sets using text- and feature-based SQL queries and scripts comparing text fields such as names or operators. Consequently, we added missing attributes like minimum and maximum allowable node pressures where available, corrected the entry and exit points based on TSO data sets, and associated the entry and exit nodes of the network with the ENTSOG's interconnection points and storage facilities. We also employed a geographical information system (GIS) to check the spatial consistency of data and complete spatial missing attributes like node height. We performed simple rule-based heuristics to check the data consistency after the augmentation.

Open gas network data provided by different sources are geospatially or temporally non-homogeneous. Similarly, the physical attributes of the network components are inconsistent in these sources. An example is the pipelines' maximum capacity values in the base data set [KKS$^+$17] and node pressure data from TSOs. The heuristic computation of maximum capacity values does not account for node pressure information. Thus, we heavily relied on S3-type preprocessing to merge data from various sources. For instance, after propagating the node pressures in the network, we again heuristically recomputed the capacities using the newly added node pressure data to improve the consistency of the data set.

Another issue with open data is its inaccuracy and incompleteness regarding active network components. So, we generated data for compressor stations, valves, and control valves. For this reason, we developed and implemented a methodology for data estimation based on partially available public data [MWSYE21]. In this method, we also considered limitations induced by the network topology. Consequently, we modeled 58 compressor stations around Germany.

We employed S3-type strategies also for preparing the supply and demand scenarios. We created a scenario generator using mathematical optimization, which uses open gas supply and demand data for Europe to generate node-based scenarios for a specific European geographical area (For the details of the scenario generator, we refer to [YEKZ24, YEKZ23].). The scenario generator consists of two interconnected LP models. The first LP model (M1) disaggregates available cumulative supply and demand data geographically using a high-level gas network optimization model of Europe gas transport system. The second one (M2) is a linear gas network optimization model of a particular European region, dispatching gas amounts found by M1 to the physical nodes of the gas network in this region. These models ensure that the gas entering and exiting a particular gas network is consistent with the capacity of the European gas transport network and the cumulative supply and demand data used as input. In this case study, we used M1 to find the gas supply and demand at cross-border interconnections and the storage facilities of the German gas network and M2 to dispatch these amounts to its nodes.

After employing the GNDP strategies above, we started analysis with the steady-state gas network optimization model (M3). The model encountered some unexpected results and infeasibility [YEMW$^+$23]. Analyzing the results, we realized that these cases might have resulted from data errors we failed to diagnose by the visualization and consistency check heuristics. We exploited knowledge from the relationship of the mathematical models as in S4-type strategies for further analysis of these results. In contrast to case study 1, we could not discuss the errors with data owners. We, therefore, employed the mathematical models M1 and M2, together with M3, for data error diagnosis.

The European high-level gas network topology data published by ENTSOG [ENT18b] is the most reliable data of this analysis. This data is input to M1. M2 uses the Ger-

man physical gas network data of the improved data set. Hence, the corrected pipeline capacities are input to M2. On the other hand, M3 uses augmented and generated data. Pipeline capacity computation is inherent to the M3.

In our analysis, we saw that the pipeline capacity attribute input to M2 was inconsistent with the pipeline capacities explicitly computed by M3. We estimated the former heuristically based on the method reported in the initial data set [KKS$^+$17] with the augmented node pressure data. Hence, realizing that the assumptions in this step were inaccurate, we implemented alternative heuristics to improve pipeline capacity estimation. We exploited the knowledge of the consistency relationship of the three models to select the correct method. Correcting the pipeline capacities in the data set also corrected the generated scenarios [YEKZ24].

As a result, we successfully compiled an admissibly accurate and complete data set of the German gas network that meets the requirements of the GasLib data model. The data set allows modeling the operational modes of the existing compressor stations thanks to the generated compressor station data. Besides conforming to the GasLib data standard, the data set can serve as a benchmark data set for stationary gas network optimization models. As a by-product, we improved the existing gas network topology data set due to [KWH$^+$17]. We made it consistent with the higher-level ENTSOG interconnection points capacity data [ENT19] and the physical network models.

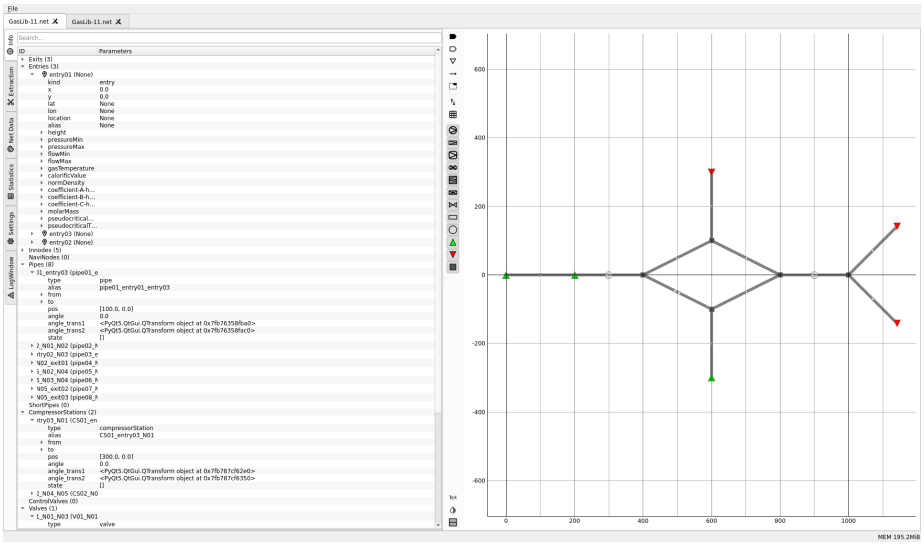
8 Summary of Case Study Results

For case study 1, we integrated data that covers the seven-year history of one of the largest German gas networks, containing approximately 12,000 km of pipelines. With this data, we built an algorithmic-intelligence-based dispatch-decision support system for our industry partner. This system helps our industry partner efficiently control their network. It is especially beneficial for uncommon cases when the experience of the operators is insufficient due to disruptive events. An example is, the change in gas direction with the political situation in Europe. Again, integrating emerging technologies, especially those related to hydrogen, is another example. Employing the compiled data, we proceed with our forecasting and optimization algorithms to achieve many valuable research goals [PCG$^+$22, HBHL$^+$21b, HAHB$^+$21].

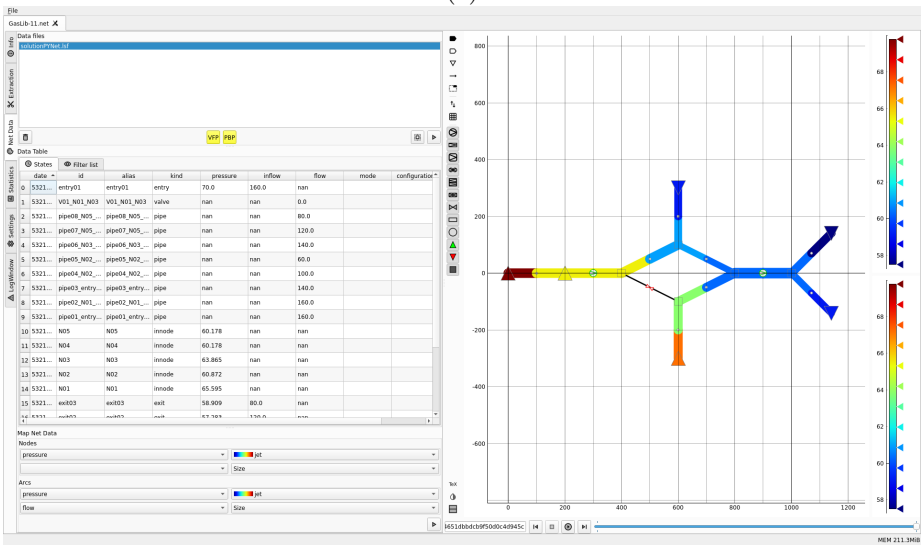
In case study 2, we compiled a German gas network data set from open data sources. Here, we employed data scattered to diverse public data sources. These sources have different formats ranging from structured database tables to unstructured web data, including pictures, lists, and specification sheets. Besides, we integrated network topology data with the supply and demand forecasts from ENTSOG. Consequently, we can explore the flexibility of the energy system provided by the gas network while researching the optimal European energy transition pathways. For instance, we employed this data set for modeling the transition of the multimodal pan-European energy system with an integrated analysis of electricity and gas transport [YEMW$^+$23] in the project plan4res [pla18].

In the case studies, we highly benefited from visualization. For the visualization of gas network topology data, we developed a tool called PyNet [Zus22]. By PyNet, we visualize the network data that is valid against the GasLib data model [Zus18a, SAB$^+$17] as in Figure 4 (a). Moreover, we visualize the results of the optimization models as in the righthand side of Figure 4 (b). In the latter case, PyNet lets us identify data errors utilizing small test scenarios. Likewise, we can observe unexpected results given real-life scenarios due to data errors. We published PyNet as an open-source software package at the repository [Zus22].

These two case studies taught us an invaluable lesson on GNDP: independent of the effort to compile a real-world data set, some data issues could remain unsolved.



(a)



(b)

Figure 4: PyNet for visualization of gas networks (**a**: GasLib-11 [SAB+17] Network Visualisation) (**b**:GasLib-11 [SAB+17] Visualization of a Feasible Solution)

Either these are structural, or the cost to adjust the corresponding source systems or transformation is unreasonably high compared to the benefit of a correction. In these cases, we can adjust data manually if errors regularly occur and the number of single issues is manageable by hand. Otherwise, we can explore means to deal with missing or erroneous data in our analysis.

For instance, we could manually adjust data in our case studies, as for the compressor machines' feasible operating ranges data in Section 6.4. Here, we also employed heuristics for initial state inconsistency to make our analysis method robust. However, in neither case study, designing algorithmic approaches that are absolutely robust against data errors was possible. Instead, we had to decide whether the data was admissibly correct for the analysis before we terminated the GNDP. Therefore, as a second lesson, we highlight the criticality of the data quality assessment ability, by which we can decide the termination of GNDP tasks.

Finally, we learned the necessity of continuously monitoring the results of the decision analysis against potential data errors. In general, we should acknowledge each data error that turns out to be too complex to fix by GNDP. First, we should try enhancing the subsequent algorithms so that they do not rely on the correctness of this type of data. If we can not succeed, we should inform the model developers and analysts about the potential decision anomalies led by data errors. Even if we do not have such a data error that we know apriori, we should scrutinize the systematically occurring infeasible solutions or unexpected results. In the case studies, our analysis encountered similar anomalies in the analysis results originating from data errors, as we later found out.

Either because of a known potential data error or the one encountered during analysis, knowledge from the data owners, as in Section 6.4, or mathematical models, as in Section 7.4 can be employed for diagnosing and correcting the error. Here, we strongly emphasize that when the data is complex and highly connected, like gas network data, even diagnosing such an error is a significant achievement.

9 Conclusion

In this paper, we acknowledge the dependence of reliable decision-making on data quality, particularly for the gas network optimization domain. However, resources dedicated to GNDP in decision-making-related projects are limited. Therefore, we propose integrating GNDP tasks into the decision-making workflow. To employ this integration, we present a data quality assessment methodology based on modeling requirements. We derive these requirements from decision needs. Besides, we group potential GNDP tools and methods in strategies with varying implementation complexity and domain knowledge requirement. Based on these strategy groups, we provide a guideline for adopting a project-specific GNDP strategy. We present two different case studies to demonstrate the implementation of the proposed workflow and strategy guide.

We address the gas network data from different angles through two case studies. One of these case studies utilizes data from a data owner and the other from public data sources. In recent years, some disruptive events caused the accumulated experience in decision-making to become obsolete for any level of decisions related to gas networks, from operational to strategic. Hence, consistent data has become even more significant. In the industrial setting, having such data increases the competition power of a TSO by enabling efficient network operation. Besides, data is essential to explore efficient ways to repurpose the network by comprehending the impacts of the changes in the commodity structure. On the other hand, research relies on open data. In contrast to the vast amount of energy system models in the literature, open data, particularly for gas transport, is a bottleneck in the European energy transition research [Ope22b].

On the other hand, publishing open gas network data is only possible when it is also allowed by the data owner. For instance, data can be confidential, as in case study 1. Likewise, data sources may have terms and conditions for data utilization preventing third parties from sharing the data even after processing it, as in the case of [CBB+09, LSS+19]. However, comparative research needs common problem instances and related data sets. Thus, in case study 2, we pay the utmost attention to preparing a data set that we can publicly share.

Generally, open-access, well-prepared, and realistic data sets are invaluable. Researchers do not like to spend much effort on cleaning-up data. As a result, good public data sets are used for decades and become benchmarks [Nat14]. Such data sets are essential to compare different modeling approaches and different scenarios. MIPLIB [GHG+21] is an excellent example of how to benchmark libraries boost research and collaboration in a specific research area.

For the pan-European gas network, benchmark data sets consistent with real-world data are scarce (please see Section 2). However, there is a high demand for such data sets to complement the energy transition research in Europe, which is substantially more advanced regarding models. We observe this need from the use of the existing data. For instance, 160 papers cited [CBB+09], although not all focus on gas networks. The LKD-EU [KKS+17] data set has about 3500 views and more than 800 downloads. GasLib gas network instances [SAB+17, Zus18a] have been cited by 140 papers. Hence, we devote the utmost attention to complementing the GasLib instances with new realistic gas network instances as presented in 7. We believe that efforts on data quality improvement of the benchmark libraries are essential in novel algorithmic decision-making solutions research to deal with future challenges. In this regard, we continue our work on data quality to design a data rating method that allows continuous data monitoring and automatic data detection and correction [YEKHZ22]. We also extend our research to supply infrastructure and utility network data other than gas networks.

Acknowledgments

The work for the results reported in Section 6 has been conducted in the Research Campus MODAL funded by the German Federal Ministry of Education and Research (BMBF) (fund numbers 05M14ZAM, 05M20ZBM).
The reported results in Section 7 received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 773897 and has been partly conducted in the Research Campus MODAL funded by the German Federal Ministry of Education and Research (BMBF) (fund numbers 05M14ZAM, 05M20ZBM)

Abbreviations

ENTSOG	European Network of Transmission System Operators for Gas
FNB	Vereinigung der Fernleitungsnetzbetreiber Gas e.V.
GIE	Gas Infrastructure Europe
GNDP	Gas Network Data Preprocessing
LP	Linear Programming
MINLP	Mixed Integer Nonlinear Programming
P2G	Power-to-Gas
PyNet	Python Network Visualization Tool
TP	Transparency Platform
TSO	Transmission System Operator

A Gas network optimization problems

An overview of the gas network optimization problem types are presented in Table 2.

References

[ACD+13] Nicola Askham, Denise Cook, Martin Doyle, Helen Fereday, Mike Gibson, Ulrich Landbeck, Rob Lee, Chris Maynard, Gary Palmer, and Julian Schwarzenbach. The six primary dimensions for data quality assessment. Technical report, DAMA United Kingdom, United Kingdom, 2013.

[AGN15] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Profiling relational data: a survey. *VLDB Journal*, 24(4):557–581, 2015. doi:10.1007/s00778-015-0389-y.

[AGS19] Florian Arnold, Michel Gendreau, and Kenneth Sörensen. Efficiently solving very large-scale routing problems. *Computers & Operations Research*, 107:32–42, 2019. URL: https://misc.sciencedirect.com/science/article/pii/S0305054819300668, doi:10.1016/j.cor.2019.03.006.

[AT22] Marcel Altendeitering and Martin Tomczyk. A Functional Taxonomy of Data Quality Tools: Insights from Science and Practice. In *Wirtschaftsinformatik 2022 Proceedings*, number February, 2022. URL: https://aisel.aisnet.org/wi2022/business_analytics/business_analytics/4.

[BBC+21] Sogol Babaeinejadsarookolaee, Adam Birchfield, Richard D. Christie, Carleton Coffrin, Christopher DeMarco, Ruisheng Diao, Michael Ferris, Stephane Fliscounakis, Scott Greene, Renke Huang, Cedric Josz, Roman Korab, Bernard Lesieutre, Jean Maeght, Terrence W. K. Mak, Daniel K. Molzahn, Thomas J. Overbye, Patrick Panciatici, Byungkwon Park, Jonathan Snodgrass, Ahmad Tbaileh, Pascal Van Hentenryck, and Ray Zimmerman. The power grid library for benchmarking ac optimal power flow algorithms. Technical report, 2021. doi:10.48550/ARXIV.1908.02788.

[BCFM09] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), 2009. doi:10.1145/1541880.1541883.

[BF18] Herib Blanco and André Faaij. A review at the role of storage in energy systems with a focus on power to gas and long-term storage. *Renewable and Sustainable Energy Reviews*, 81:1049–1086, 2018. URL: https://misc.sciencedirect.com/science/article/pii/S1364032117311310, doi:10.1016/j.rser.2017.07.062.

[BGH+19] Peter Benner, Sara Grundel, Christian Himpe, Christoph Huck, Tom Streubel, and Caren Tischendorf. *Gas Network Benchmark Models*, pages 171–197. Springer International Publishing, 2019. doi:10.1007/11221_2018_5.

[BGK+14] Andreas Betker, Inken Gamrath, Dirk Kosiankowski, Christoph Lange, Heiko Lehmann, Frank Pfeuffer, Felix Simon, and Axel Werner. Comprehensive topology and traffic model of a nationwide telecommunication network. *J. Opt. Commun. Netw.*, 6(11):1038–1047, Nov 2014. URL: https://opg.optica.org/jocn/abstract.cfm?URI=jocn-6-11-1038, doi:10.1364/JOCN.6.001038.

Table 2: Examples of gas transport-related decisions and model requirements

Decision Type	Example Decision	Modeling Detail	Computational Size		Modeling Precision	Typical Models
			Geographical	Temporal		
Strategic	Required cumulative pipeline capacity between countries/gas systems to maintain security of supply of the continent under political requirements	Flow between countries / important gas systems	Span: Continent Res: Countries or Gas systems	Span: Decades/years	LP	High-level gas network optimization model
Tactical	A routing of gas in the gas pipeline network given supply and demand	Flow of gas on physical gas network consisting of pipelines	Span: Country Res: Physical nodes of gas network	Span: A year/months	LP	Linear gas network optimization model
	A steady-state routing of gas in the network and resulting gas pressure at the nodes given supply and demand	Steady-state flow of gas on physical gas network consisting of pipelines, pressure of gas at the nodes of the network	Span: Country Res: Physical nodes of gas network	Span: A year/months	NLP	(Quasi-)Steady state gas network optimization model
	A steady-state routing of gas in the network, resulting gas pressure at the nodes, and operational state of the complex facilities of the network	Steady-state flow of gas on physical gas network consisting of pipelines, pressure of gas at the nodes of the network, operational state of the complex facilities on gas network	Span: Country Res: Physical nodes of gas network	Span: A year/ months	MINLP	Stationary gas network optimization model
Operational	An operational setting of the network to route forecasted amount of gas with the least amount of change in the operational state of complex facilities	Time dependent flow of gas on physical gas network consisting of pipelines, pressure of gas at the nodes of the network, operational state of the complex facilities on gas network	Span: Subcountry network Res: Physical nodes of gas network	Span: Days/ Hours Res: Hours/ Minutes	MINLP	Transient gas network optimization model

[BGR+15] Felix Biessmann, Jacek Golebiowski, Tammo Rukat, Dustin Lange, and Philipp Schmidt. Automated Data Validation in Machine Learning Systems. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2015.

[BMW23] BMWK. Power-to-gas, 2023. Last accessed 9 February 2023. URL: `https://misc.bmwk.de/Redaktion/EN/Artikel/Energy/gas-power-to-gas.html`.

[BS20] Carlo Batini and Monica Scannapieco. *Data and Information Quality: Dimensions, Principles and Techniques*. Springer, 2020. `doi:10.1201/9781482264654-14`.

[CBB+09] Rui Carvalho, Lubos Buzna, Flavio Bono, Eugenio Gutiérrez, Wolfram Just, and David Arrowsmith. Robustness of trans-european gas networks. *Phys. Rev. E*, 80:016106, Jul 2009. URL: `https://link.aps.org/doi/10.1103/PhysRevE.80.016106`, `doi:10.1103/PhysRevE.80.016106`.

[CBvAO+21] Sandrine Charousset-Brignol, Wim van Ackooij, Nadia Oudjane, Dominique Daniel, Slimane Noceir, Utz-Uwe Haus, Alfio Lazzaro, Antonio Frangioni, Rafael Lobato, Ali Ghezelsoflu, Niccolò Iardella, Laura Galli, Enrico Gorgone, Mauro dell'Amico, Spyros Giannelos, Alex Moreira, Goran Strbac, Stefan Borozan, Paula Falugi, Danny Pudjianto, Lothar Wyrwoll, Carlo Schmitt, Marco Franken, Daniel Beulertz, Henrik Schwaeppe, Dieter Most, Inci Yüksel-Ergün, Janina Zittel, and Thorsten Koch. Synergistic approach of multi-energy models for a european optimal energy system management tool. *The project repository journal*, 9:113 – 116, 2021.

[CZ15] Li Cai and Yangyong Zhu. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14:1–10, 2015. `doi:10.5334/dsj-2015-002`.

[DWS00] Daniel De Wolf and Yves Smeers. The gas transmission problem solved by an extension of the simplex algorithm. *Management Science*, 46(11):1454–1465, 2000.

[EEE18] ENTSOG and ENTSO-E. Power to gas-a sector coupling perspective, 2018. Last accessed 10 October 2022. URL: `https://misc.entsoe.eu/2018/10/15/power-to-gas-a-sector-coupling-perspective/`.

[ENT18a] ENTSOG. Transmission Capacity and System Development Maps, 2018. Last accessed 17 June 2022. URL: `https://misc.entsog.eu/maps#system-development-map`.

[ENT18b] ENTSOG. Transparency platform, 2018. Last accessed 10 October 2022. URL: `https://transparency.entsog.eu/`.

[ENT18c] ENTSOG. Web site, 2018. Last accessed 10 October 2022. URL: `https://entsog.eu/`.

[ENT19] ENTSOG. ENTSOG TYNDP 2018 – Annex C - Capacities, 2019. Last accessed 17 June 2022. URL: `https://misc.entsog.eu/sites/default/files/2019-05/TYNDP2018-AnnexC-Capacities.xlsx`.

[Eur09] European Parliament and the Council of the European Union. Regulation (ec) no 715/2009 of the european parliament and of the council on conditions for access to the natural gas transmission networks, 2009. Last accessed 10 October 2022. URL: `https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:211:0036:0054:en:PDF`.

[EW22] Lisa Ehrlinger and Wolfram Wöß. A Survey of Data Quality Measurement and Monitoring Tools. *Frontiers in Big Data*, 5(March), 2022. doi:10.3389/fdata.2022.850611.

[FAO+20] F. Fossatti, G. Agugiaro, L. Olde Scholtenhuis, A. Dorée, and F. Fossatti. Data Modeling for Operation and Maintenance of Utility Networks: Implementation and Testing. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 6(4/W1):69–76, 2020. doi:10.5194/isprs-annals-VI-4-W1-2020-69-2020.

[FCSB03] Craig W. Fisher, InduShobha Chengalur-Smith, and Donald P. Ballou. The impact of experience and time on the use of data quality information in decision making. *Information Systems Research*, 14(2):170–188, 2003. doi:10.1287/isre.14.2.170.16017.

[FG18] FNB-Gas. Fnb-gas network development plan data repository, 2018. Last accessed 10 October 2022. URL: https://misc.nep-gas-datenbank.de:8080/app/#\protect\protect\leavevmode@ifvmode\kern-.1667em\relax/.

[FGG+15] Armin Fügenschuh, Björn Geißler, Ralf Gollmer, Antonio Morsi, Marc E. Pfetsch, Jessica Rövekamp, Martin Schmidt, Klaus Spreckelsen, and Marc C. Steinbach. *Chapter 2: Physical and technical fundamentals of gas networks*, pages 17–43. PA, Phyledelpia, 2015. doi:10.1137/1.9781611973693.ch2.

[Flu22] Fluxys Germany. Electronic data platform for transmission, 2022. Last accessed 15 December 2022. URL: https://gasdata.tnp.gsmartsuite.com/en/transmission/.

[Gas21a] Gas Infrastructure Europe (GIE). Gie aggregated gas storage inventory - storage data, 2021. Last accessed 24 September 2021. URL: https://agsi.gie.eu/#/.

[Gas21b] Gas Infrastructure Europe (GIE). Gie web site, 2021. Last accessed 24 September 2021. URL: https://misc.gie.eu/.

[Gas22] Gascade. Network data, 2022. Last accessed 15 December 2022. URL: https://tron.gascade.biz/?language=en#.

[GHG+21] Ambros Gleixner, Gregor Hendel, Gerald Gamrath, Tobias Achterberg, Michael Bastubbe, Timo Berthold, Philipp M. Christophel, Kati Jarck, Thorsten Koch, Jeff Linderoth, Marco Lübbecke, Hans D. Mittelmann, Derya Ozyurt, Ted K. Ralphs, Domenico Salvagnin, and Yuji Shinano. MIPLIB 2017: Data-Driven Compilation of the 6th Mixed-Integer Programming Library. *Mathematical Programming Computation*, 2021. doi:10.1007/s12532-020-00194-3.

[GRT22] GRTGaz Deutschland. Megal pipeline system technical parameters, 2022. Last accessed 15 December 2022. URL: https://misc.grtgaz-deutschland.de/infrastructure/.

[GTGnd] GTG Nord. Transparency platform, n.d. Last accessed 15 December 2022. URL: https://b2b-prod.gtg-nord.de/publication/#.

[HAHB+21] Felix Hennings, Lovis Anderson, Kai Hoppmann-Baum, Mark Turner, and Thorsten Koch. Controlling transient gas flow in real-world pipeline intersection areas. *Optimization and Engineering*, 22:687 – 734, 2021. doi:10.1007/s11081-020-09559-y.

[HBHL+21a] Kai Hoppmann-Baum, Felix Hennings, Ralf Lenz, Uwe Gotzes, Nina Heinecke, Klaus Spreckelsen, and Thorsten Koch. Optimal Operation of Transient Gas Transport Networks. *Optimization and Engineering*, 22(2):735–781, 2021. doi:10.1007/s11081-020-09584-x.

[HBHL+21b] Kai Hoppmann-Baum, Felix Hennings, Ralf Lenz, Uwe Gotzes, Nina Heinecke, Klaus Spreckelsen, and Thorsten Koch. Optimal operation of transient gas transport networks. *Optimization and Engineering*, 22:735 – 781, 2021. doi:10.1007/s11081-020-09584-x.

[HHH+15] Benjamin Hiller, Christine Hayn, Holger Heitsch, René Henrion, Hernan Leövey, Andris Möller, and Werner Römisch. *Chapter 14: Methods for verifying booked capacities*, pages 291–315. PA, Phyledelpia, 2015. doi:10.1137/1.9781611973693.ch14.

[HHK+18] Bernd Heinrich, Diana Hristova, Mathias Klier, Alexander Schiller, and Michael Szubartowicz. Requirements for data quality metrics. *Journal of Data and Information Quality*, 9(2), 2018. doi:10.1145/3148238.

[HN20] Mazhar Hameed and Felix Naumann. Data Preparation: A Survey of Commercial Tools. *SIGMOD Record*, 49(3):18–29, 2020. doi:10.1145/3444831.3444835.

[HSW21] Benjamin Hiller, René Saitenmacher, and Tom Walther. Improved models for operation modes of complex compressor stations. *Mathematical Methods of Operations Research*, 94(2):171–195, 2021.

[HW18] Benjamin Hiller and Tom Walther. Modelling compressor stations in gas networks. Technical Report 17-67, ZIB, Takustr. 7, 14195 Berlin, 2018.

[IN22] Ihab F. Ilyas and Felix Naumann. Data Errors: Symptoms, Causes and Origins. *Bulletin of the Technical Committee on Data Engineering*, 45(1):4–9, 2022. URL: http://sites.computer.org/debull/A22mar/issue1.htm.

[JBE+20] Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, and Tomasz Janowski. Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly*, 37(3):101493, 2020. doi:10.1016/j.giq.2020.101493.

[JSSW15] I. Joormann, M. Schmidt, M.C. Steinbach, and B.M. Willert. What does feasible mean? In Thorsten Koch, Benjamin Hiller, M. E. Pfetsch, and Lars Schewe, editors, *Evaluating Gas Network Capacities*, chapter 11, pages 211–232. SIAM-MOS Series on Optimization, PA, Phyledelpia, 2015.

[KDS12] Yuliya V. Karpievitch, Alan R. Dabney, and Richard D. Smith. Normalization and missing value imputation for label-free lc-ms analysis. *BMC Bioinformatics*, 13(16):S5, November 2012. URL: https://doi.org/10.1186/1471-2105-13-S16-S5.

[KHPS15] Thorsten Koch, Benjamin Hiller, M. E. Pfetsch, and Lars Schewe, editors. *Evaluating Gas Network Capacities*. SIAM-MOS Series on Optimization, PA, Phyledelpia, 2015.

[KKP07] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas. Data preprocessing for supervised leaning. *International Journal of Computer and Information Engineering*, 1(12):4104 – 4109, 2007.

[KKS+17] Friedrich Kunz, Mario Kendziorski, Wolf-Peter Schill, Jens Weibezahn, Jan Zepter, Christian von Hirschhausen, Philip Hauser, Matthias Zech, Dominik Moest, Sina Heidari, Björn Felten, and Christoph Weber. Data Documentation 92: Electricity, Heat, and Gas Sector Data for Modeling the German System, 2017. Last accessed 15 July 2024. URL: https://d-nb.info/1154864928/34.

[KWH+17] Friedrich Kunz, Jens Weibezahn, Philip Hauser, Sina Heidari, Wolf-Peter Schill, Björn Felten, and Christoph Weber. Reference Data Set: Electricity, Heat, and Gas Sector Data for Modeling the German System (Version 1.0.0), 2017. URL: https://zenodo.org/records/1044463, doi:10.5281/zenodo.1044463.

[LE18] LKD-EU. Lkd-eu project website, 2018. Last accessed 16 September 2021. URL: https://misc.diw.de/de/diw_01.c.537097.de/projekte/langfristige_planung_und_kurzfristige_optimierung_des_elektrizitaetssystems_in_deutschland_im_europaeischen_kontext_lkd_eu.html.

[LHGM22] Lukas Löhr, Raphael Houben, Carolin Guntermann, and Albert Moser. Nested decomposition approach for dispatch optimization of large-scale, integrated electricity, methane and hydrogen infrastructures. *Energies*, 15(8), 2022. URL: https://misc.mdpi.com/1996-1073/15/8/2716, doi:10.3390/en15082716.

[Li12] B. Li. *Simulation and capacity calculation in real German and European interconnected gas transport systems*. Cuvillier Verlag, Goettingen, 2012.

[Loc21] Stevie Lochran. Gnome: A dynamic dispatch and investment optimisation model of the european natural gas network and its suppliers. In *Operations Research Forum*, volume 2, pages 1–44. Springer, 2021.

[LSS+19] Peter Lustenberger, Felix Schumacher, Matteo Spada, Peter Burgherr, and Bozidar Stojadinovic. Assessing the performance of the European natural gas network for selected supply disruption scenarios using open-source information. *Energies*, 12(24), 2019. doi:10.3390/en12244685.

[MDR15] Michael J. Mortenson, Neil F. Doherty, and Stewart Robinson. Operational research from Taylorism to Terabytes: A research agenda for the analytics age. *European Journal of Operational Research*, 241(3):583–595, 2015. URL: http://dx.doi.org/10.1016/j.ejor.2014.08.029, doi:10.1016/j.ejor.2014.08.029.

[MGYE+20] Dieter Most, Spyros Giannelos, Inci Yueksel-Erguen, Daniel Beulertz, Utz-Uwe Haus, Sandrine Charousset-Brignol, and Antonio Frangioni. A novel modular optimization framework for modelling investment and operation of energy systems at european level. Technical Report 20-08, ZIB, Takustr. 7, 14195 Berlin, 2020.

[MWSYE21] Dieter Most, Lothar Wyrwoll, Carlo Schmitt, and Inci Yueksel-Erguen. plan4res D2.2 - Case Study 1 Report - Multimodal energy concept for achieving Europe's carbon reduction goals. Technical report, 2021. doi:10.5281/zenodo.5809338.

[Nat14] National Institute of Standards and Technology. Nist handprinted forms and characters database, 2014. accessed on 10.10.2022. URL: http://doi.org/10.18434/T4H01C.

[OGP06] Matthew Eric Otey, Amol Ghoting, and Srinivasan Parthasarathy. Fast distributed outlier detection in mixed-attribute data sets. *Data Mining and Knowledge Discovery*, 12(2):203–228, May 2006. URL: https://doi.org/10.1007/s10618-005-0014-6.

[ONT22] ONTRAS. Transparency data, 2022. Last accessed 15 December 2022. URL: https://portal.ontras.com/portal.public/transparency.

[Ope17] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org , 2017. URL: https://misc.openstreetmap.org.

[Ope22a] Open Energy Platform. Open energy platform, 2022. Last accessed 19 January 2023. URL: https://openenergy-platform.org/.

[Ope22b] Open Energy Platform. Open energy platform web site, 2022. Last accessed 24 November 2022. URL: https://openenergy-platform.org.

[Ope22c] Open Grid Europe (OGE). Network data web publications, 2022. Last accessed 15 December 2022. URL: https://oge.net/en/for-customers/gas-transmission/market-information/operational-network-data/web-publications.

[OWPT10] S. Orlowski, R. Wessäly, M. Pióro, and A. Tomaszewski. Sndlib 1.0—survivable network design library. *Networks*, 55(3):276–286, 2010. doi:10.1002/net.20371.

[PCG+22] Milena Petkovic, Ying Chen, Inken Gamrath, Uwe Gotzes, Natalia Selini Hadjidimitrou, Janina Zittel, Xiaofei Xu, and Thorsten Koch. A hybrid approach for high precision prediction of gas flows. *Energy Systems*, 13:383 – 408, 2022. doi:10.1007/s12667-021-00466-4.

[PGH+15] Marc E Pfetsch, Ralf Gollmer, Benjamin Hiller, Jesco Humpola, Thorsten Koch, Thomas Lehmann, Alexander Martin, Antonio Morsi, Lars Schewe, Martin Schmidt, Robert Schwarz, Jonas Schweiger, Claudia Stangl, Marc C Steinbach, Stefan Vigerske, and Bernhard M Willert. Validation of nominations in gas network optimization: models, methods, and solutions. *Optimization Methods and Software*, 30(1):15–53, 2015.

[pla18] plan4res. plan4res project website, 2018. Last accessed 16 September 2021. URL: https://misc.plan4res.eu/.

[PS08] Rosanne Price and Graeme Shanks. Data quality and decision making. In *Handbook on Decision Support Systems 1*, pages 65–82. Springer, 2008.

[QSP+22] Marco Quagliotti, Laura Serra, Annachiara Pagano, Panos Groumas, Paraskevas Bakopoulos, and Hercules Avramopoulos. Disaggregation and cloudification of metropolitan area networks: Enabling technologies and impact on architecture, cost, and power consumption [Invited]. *Journal of Optical Communications and Networking*, 14(6):C38–C49, 2022. doi:10.1364/JOCN.450822.

[RKCMP19] Laura Rettig, Mourad Khayati, Philippe Cudré-Mauroux, and Michał Piorkówski. *Online Anomaly Detection over Big Data Streams*, pages 289–312. Springer International Publishing, 2019. doi:10.1007/978-3-030-11821-1_16.

[RKD+13] Franz Rambach, Beate Konrad, Lars Dembeck, Ulrich Gebhard, Matthias Gunkel, Marco Quagliotti, Laura Serra, and Victor López. A multilayer cost model for metro/core networks. *Journal of Optical Communications and Networking*, 5(3):210–225, 2013. doi:10.1364/JOCN.5.000210.

[RMS+22] Hendrik Roth, Simon Paul Mönch, Thomas Schäffer, Hochschule Heilbronn, Simon Mönch, and Thomas Schäffer. Towards Augmented MDM: Overview of Design and Function Areas – A Literature Review. In *AMCIS 2022 Proceedings*, 2022.

[Roe14] J. Roevekamp. *Transportnetzberechnung zur Feststellung der Erdgasversorgungssicherheit in Deutschland unter regulatorischem Einfluss*. Dissertation, Technischen Universität Clausthal, Niedersachsen, 2014.

[SAB+17] Martin Schmidt, Denis Aßmann, Robert Burlacu, Jesco Humpola, Imke Joormann, Nikolaos Kanelakis, Thorsten Koch, Djamal Oucherif, Marc E. Pfetsch, Lars Schewe, Robert Schwarz, and Mathias Sirvent. Gaslib—a library of gas network instances. *Data*, 2(4), 2017. URL: https://misc.mdpi.com/2306-5729/2/4/40, doi:10.3390/data2040040.

[Sci18] SciGrid Gas Project. Scigrid gas project website, 2018. Last accessed 10 October 2022. URL: https://misc.gas.scigrid.de.

[SGR+15] Sebastian Schiebahn, Thomas Grube, Martin Robinius, Vanessa Tietze, Bhunesh Kumar, and Detlef Stolten. Power to gas: Technological overview, systems analysis and economic assessment for a case study in germany. *International Journal of Hydrogen Energy*, 40(12):4285–4294, 2015. URL: https://misc.sciencedirect.com/science/article/pii/S0360319915001913, doi:10.1016/j.ijhydene.2015.01.123.

[SLS+21] R. Schwarz, F. Lacalandra, L. Schewe, A. Bettinelli, D. Vigo, A. Bischi, T. Parriani, E. Martelli, K. Vuik, R. Lenz, H. Madsen, I. Blanco, D. Guericke, I. Yüksel-Ergün, and J. Zittel. *Network and Storage*, pages 89–105. Springer International Publishing, Cham, 2021. doi:10.1007/978-3-030-57442-0_6.

[SLW97] Diane M. Strong, Yang W. Lee, and Richard Y. Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997.

[Ter22] Terranets bw. Network data, 2022. Last accessed 15 December 2022. URL: https://misc.terranets-bw.de/en/for-customers/gas-network/market-information/network-data.

[The21] The European Hydrogen Backbone Initiative. The european hydrogen backbone (ehb) initiative website, 2021. Last accessed 14 July 2022. URL: https://ehb.eu/.

[Thy22] Thyssengas. Implementing of transparency requirements as per eg vo 715/2009, 2022. Last accessed 15 December 2022. URL: https://thyssengas.com/en/network-enquiries/transparency-information/overview.html.

[VIV+12] Ward Van Heddeghem, Filip Idzikowski, Willem Vereecken, Didier Colle, Mario Pickavet, and Piet Demeester. Power consumption modeling in optical multilayer networks. *Photonic Network Communications*, 24(2):86–102, 2012. doi:10.1007/s11107-011-0370-7.

[WOB14] Philip Woodall, Martin Oberhofer, and Alexander Borek. A classification of data quality assessment and improvement methods. *International Journal of Information Quality*, 3(4):298–321, 2014. doi:10.1504/IJIQ.2014.068656.

[WS96] Richard Y. Wang and Diane M. Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4):5–33, 1996.

[WSF95] Richard Y. Wang, Veda C. Storey, and Christopher P. Firth. A Framework for Analysis of Data Quality Research. *IEEE Transactions on Knowledge and Data Engineering*, 7(4):623–640, 1995. doi:10.1109/69.404034.

[YEKHZ22] Inci Yueksel-Erguen, Thorsten Koch, Felix Hennings, and Janina Zittel. Improving data quality in the presence of superhuman complexity in data errors. In Katsuki Fujisawa, Shizuo Kaji, Toru Ishihara, Masaaki Kondo, Yuji Shinano, Takuji Tanigawa,

and Naoko Nakayama, editors, *Construction of Mathematical Basis for Realizing Data Rating Service*, volume 90, 2022. URL: https://joint.imi.kyushu-u.ac.jp/wp-content/uploads/2022/12/29d12d46a22d702596713780acde0689-3.pdf.

[YEKZ23] Inci Yueksel-Erguen, Thorsten Koch, and Janina Zittel. Consistent flow scenario generation based on open data for operational analysis of european gas transport networks. In *Operations Research Proceedings*, 2023. accepted for publication.

[YEKZ24] Inci Yueksel-Erguen, Thorsten Koch, and Janina Zittel. Mathematical optimization based flow scenario generation for operational analysis of european gas transport networks based on open data. Technical Report 24-03, ZIB, Takustr. 7, 14195 Berlin, 2024.

[YEMW+23] Inci Yueksel-Erguen, Dieter Most, Lothar Wyrwoll, Carlo Schmitt, and Janina Zittel. Modeling the transition of the multimodal pan-european energy system including an integrated analysis of electricity and gas transport. *Energy Systems*, pages 1–46, 2023.

[Zus18a] Zuse Institute Berlin. Gaslib - a library of gas network instances, 2018. Last accessed 10 October 2022. URL: https://gaslib.zib.de.

[Zus18b] Zuse Institute Berlin. Research campus modal website, 2018. Last accessed 10 October 2022. URL: https://forschungscampus-modal.de/about-us/energy-lab-en?lang=en.

[Zus22] Zuse Institute Berlin. Pynet (python network visualization tool). https://git.zib.de/energy-public/PyNet, 2022. Last accessed 15 July 2024.