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 (September 2020)

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*

September 25, 2020

Abstract

The German high-pressure natural gas transport network consists of thousands of interconnected elements spread over more than 120,000 km of pipelines built during the last 100 years. During the last decade, we have spent many person-years to extract consistent data out of the available sources, both public and private. Based on two case studies, we present some of the challenges we encountered. Preparing consistent, high-quality data is surprisingly hard, and the effort necessary can hardly be overestimated. Thus, it is particularly important to decide which strategy regarding data curation to adopt. Which precision of the data is necessary? When is it more efficient to work with data that is just sufficiently correct on average? In the case studies we describe our experiences and the strategies we adopted to deal with the obstacles and to minimize future effort. Finally, we would like to emphasize that well-compiled data sets, publicly available for research purposes, provide the grounds for building innovative algorithmic solutions to the challenges of the future.

Keywords— GasLib, gas transport, networks, stationary gas network optimization, transient gas network optimization, real-world data consistency, company data usage, public gas network data

1 Introduction

The liberalization of the European gas market in 2009 [1] posed novel and challenging problems to transmission network operators, hereafter referred to as TSOs. Capacity planning, for example, got much more complex because the TSO has no influence anymore on where gas is supplied or stored. As another example, the 24/7 operation of the network has also become more demanding as capacity trading became much more short-term. The TSO needs to operate the system most efficiently, while at the same time maximizing capacities to offer. To guarantee the security of the supply and to meet all demands, sophisticated IT solutions are needed to support the dispatchers

at the network control center. However, advanced data-driven AI decision support systems crucially depend on reliable, coherent, and consolidated data.

An algorithmic intelligence-based support system requires data on the network topology, capacities, technical details of active elements such as compressors, as well as data from the legal contract system. The available data must comprehensively describe the current state, support forecasting, or sampling techniques by sufficiently representing the past. Finally, to allow for prescriptive analytics, the potential options for controlling the network must be described in detail.

Publicly available data sources are scarce. *The European Network of Transmission System Operators for Gas* (ENTSO-G) [2] provides regular information on gas supply and demands. However, gas network topology data is only available on PDF maps, i.e., the ENTSO-G Transmission Capacity Map [3] and the ENTSO-G/GIE System Development Map [4]. As a result of the ForNe project [5], technical gas network descriptions and gas supply and demands based on contracts are available in the GasLib collection [6, 7]. The library also includes data for realistic gas networks based on real-world network data from industry partners. However, these data had to be modified, and, in certain details, simplified, from the original. A public data set for the German high-pressure gas transport network, which alone consists of more than 120,000 km of pipelines built during the last 100 years [8, 9], was compiled within the research project 'LKD-EU' (Long-term planning and short-term optimization of the German electricity system within the European framework: Further development of methods and models to analyze the electricity system including the heat and gas sector) [10]. Although this data set includes network topology with geographical location data, it is not complete regarding the compressor station and pipeline data. These are especially significant for physical-flow based gas network optimization models. More recently, SciGridGas [11] aims at the development of methods to create an automated network model of the European gas transportation network.

The compilation of comprehensive research data sets from real-world data is often complicated. There are several reasons. In some cases, the original design of the data was done for very different purposes [12]. Sometimes, they are compiled in separate systems without any links between them [13]. Often data systems have grown historically without a joint plan from the beginning [14].

When creating real-world data for research and development purposes, a lot of preprocessing is needed. The different approaches can be categorized as following:

S1. The most common approach in mathematics is to employ the data as given with small interventions, e.g., outlier filtering [15, 16], deleting apparent errors, and replacing missing values by substitute value formation [17, 15, 18]. Once a data-set has been used by the first group, later ones ask no further questions about how realistic the data is.

S2. Another purpose of preprocessing is to anticipate problems in the data and include procedures to deal with data inconsistencies before they even arrive [19]. This strategy is important for systems that deal with online data.

S3. One option is to report the errors or inconsistencies detected during research and correct them at the source system level. To employ this strategy, the data producer needs to be involved in the project. For certain types of very complex data problems, this is the only acceptable approach to end up with useful results. Imagine, for example, a model with data that should be feasible, but an NP-hard computation is necessary to prove this. Data that can be corrected in such an iterative approach is limited. Furthermore, it is typically beyond the scope and budget of the research and development project.

S4. Finally, for data that is not available at all, or where the effort of compiling it is beyond the project scope, educated assumptions and heuristics come into consideration [17, 21, 20, 14, 22]. However, these need to be tailored for individual

research and development purposes to ensure reliable end products.

Focusing on providing data for natural gas transport solutions, we present two case studies that employ the entire range of these preprocessing methods. In Example 1, we provide information on the specific challenges from a project with the support of the data producer. Here, we need to build a robust system that will work online. From the other end of the range, Example 2 provides insights from a project relying only on data that is publicly available. Due to the scarcity of the available data, we make extensive use of S4.

Besides, the two examples cover a wide range of applications in the gas transport industry, from short-term operations to long-term capacity planning. Example 1, from the GasLab of the research campus MODAL, aims at the development of a decision support system for gas network operators [23]. The system consists of a forecasting unit, to predict the hourly supply and demand [24] as well as an optimization unit to determine the control schedule for the individual network elements with sufficient accuracy [25, 26]. Example 2, from the Horizon 2020 [27] project "Synergistic Approach of Multi-Energy Models for an European Optimal Energy System Management Tool" (plan4res) [28], focuses on the multi-energy system of Europe in the upcoming decades. By integrating the gas grid into the multi-energy system, we want to understand the restrictions and flexibilities of power-to-gas (P2G) and gas-to-power (G2P) technologies in the upcoming decades [29]. A common focus of two case studies is data processing for compressors, which needs to account for the different research goals.

In this paper, we present how the different data preprocessing strategies are used to cope with the diverse challenges in the two case studies. We provide some guidance on how to deal with the most common obstacles. In Section 3, we outline the importance of realistic gas network data with sufficient detail in our research on gas networks involving both operational problems for short-term decision support and strategic problems involving longer-term decisions. Sections 4 and 5 describe the data preprocessing in the two case studies. We explain the problems we encountered and give reasons why they occurred. We provide notes on the strategies to deal with them. 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 a large amount of effort.

2 Related Work

Research concerning big data can, in general, be divided into big data management (BDM) and big data analytics (BDA) [30]. While BDM focusses on the collection, preprocessing, storage, and sharing of the data, BDA investigates different ways of using the data to gain knowledge. Industrial applications of these two in the energy domain can, for example, be found in [30] and [31], respectively. As mentioned above, in this paper we present different data preprocessing strategies and measures that we used when dealing with data concerning gas transportation networks. Hence, we do not focus on the storage and accessibility of the corresponding database, as for example investigated by [32]. Instead, we focus on identifying and resolving data problems that occurred during the collecting of the data and explicitly exclude the trust-building in the data storage system itself, which was, for example, discussed in [33] or [34]. Regarding the preprocessing approaches, we used methods of the four strategies introduced above, where we also listed relevant literature concerning these.

The studies on collecting public gas transport data to build a data set are either holistic studies that capture both the economic and physical features of the gas transport system [35, 11, 22] or focus only on the physical features [20, 17], i.e., the network topology. The former is holistic in the sense that they can be used in the end-to-end analysis of gas transport networks starting from the demand and supply and resulting

in the flow of gas through the network. Among those, [35, 22] considers the German gas network infrastructure, whereas [22] claims that their methodology can be extended to the European gas transport network, and [11] aims to build a data set for European gas transport network. The latter builds benchmark network topology data only. [17] uses predefined data from a data supplier and focuses on missing parts of EU network topology. They used methods for substituting the missing data, like for example pipeline capacities. [20] uses heuristics and assumptions based on gas network mathematical modeling knowledge to build benchmark gas network data sets for research. Besides, there are studies in between such as [36], which includes economic data as well as network components data such as compressor stations of the German gas transport network, though the complete network topology data is not included. On the other hand, there are studies that focus on the demand and supply part of the data such as [21], which uses heuristics to generate adversarial nomination data for the gas network optimization models to apply a sort of a stress test to the network.

3 Requirements on the Data

To provide solutions to any real-world problem, realistic data is essential that captures the actual properties of the entities under consideration as closely as possible. Collecting this data and ensuring its thorough quality and consistency often requires a significant amount of effort. Consequently, it is essential to specify the most relevant data for the considered problem as well as the quality of that data that is necessary to solve it.

We deal with gas network optimization problems for both, long-term planning and short term operation. In long-term planning, our main focus is to investigate whether the existing capacity of the gas network is sufficient for transporting supplied gas to the gas-demanding regions. Here, the geographical scale of scenarios is typically country- or continent-level, and the time span varies from years to decades. In short-term planning, our focus are operational decisions such as compressor station configurations and opened/closed control valves that enable the realization of the gas transport demands. The time resolution for short-term planning problems vary from minutes to hours. Thus the focus of the data preparation processes differs concerning the geographical scale, temporal span, and time-resolution depending on whether we are investigating long-term planning or short-term operational decisions.

One example for the change in the focus is the accuracy of the compressor station models. For long-term operation, we are mainly interested in the maximal compression capabilities, while short-term operations need more detailed models, which take the individual internal machinery into account and consequently need more data to be described. The data needed for this more detailed description of a compressor station will be presented below. On the other hand, long-term planning problems require topological data relevant to a larger geographical area, which shifts the focus of the data collection to consistently merging different data sets belonging to various TSOs, and accurate forecasting the supply and demand for upcoming years to decades at country- or EU-level.

While the above-given descriptions are valid for any problem concerning large scale network infrastructure, dealing with gas network data is a particular challenge. The transport of gas is facilitated by increasing or decreasing the pressure at certain points in the network. These pressure differences induce flow that travels from high-pressure areas to low-pressure areas. As a consequence, the maximal amount of gas flow over a particular pipeline, which is often just referred to as *capacity*, depends on the size of the maximal pressure differences between its end nodes. This quantity depends on a multitude of different factors, like the technical pressure bounds and velocity limits of the pipeline itself or nearby network elements, the capabilities of the

compressors present in the area, but also their current availability. Hence, this capacity of a pipeline or an even bigger part of the network can only be determined using a detailed representation of the network topology as well as an at least halfway accurate description all the technical properties and logical dependencies of all the involved elements, see for example the work of Koch et al. [37]. High-level models and/or unrealistic, assumption-based data lacking a sufficient level of detail, i.e., estimated flow capacities on pipelines, result in mathematical problems that tend to be either infeasible or yield a trivial solution.

One strategy to improve the quality of a given data set is to gain insights while working with it, i.e., by solving optimization models parameterized by the data. Unexpected model behavior can then identify data inconsistencies and errors. However, the above discussion shows that the main restrictions of the network are related to limiting factors regarding the pressure, which may result from the complex interaction of different network elements including compressor stations and control valves. For this reason, the source of the infeasibility might involve a large number of network elements and therefore parameters. Thus, finding the one causal data error from this vast amount is most often a challenge.

As an example of the data needed to optimize the operation of gas networks, we have a closer look at compressor stations. The description is based on [38], which provide even more background information on the topic and explicit formulas for all mentioned relationships. Compressor stations are used to increase the pressure of the gas along the direction of the flow. To do so, each one uses a set of corresponding compressor units. Each unit represents the combination of a single compressor machine, the actual technical element compressing the gas, and a corresponding drive, providing the power needed for the machine. The compressor units can be arranged in series or parallel to allow for a higher compression ratio or throughput in terms of gas flow. The predefined set of all technically possible arrangements is called the set of configurations of a compressor station.

For each of the single compressor machines, we have a feasible range in which it can be operated in and which depends on the corresponding type of the compressor machine. Furthermore, each point in this feasible operating range is associated with an efficiency value influencing the power needed to compress gas at the corresponding conditions. The parameters describing the feasible operating range are either given as technical properties of the compressor machine or have to be fitted based on measured values. For the drives, there are again different types used in practice. Here, the relevant quantities are the energy consumed by the drive to provide a corresponding amount of power and the maximum power that can be provided in general. The maximum power depends on the compressor's current speed as well as the temperature of the ambient air that is used for cooling.

To summarize, we need to know for each compressor station the corresponding set of compressor units, their possible arrangements given as the 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.

4 Example 1: Company data from Gas Transmission System Operators

In the GasLab project of the Research Campus MODAL, we develop a decision support system for gas network operators in close cooperation with our industry partner. In the following section, we present the preprocessing measures we applied to the data to ensure a stable operation of the system. Therefore, we first describe which data is required in general, present the data interface we created, describe the data problems we encountered, and the solutions we found for them.

4.1 Data requirements

From a research point of view, a consistent, complete, and longtime spanning data set in one fixed format would enable us to train our gas flow forecasting system seamlessly, validate the optimization algorithms finding future control measures against a multitude of different network situations occurring over time, and maintain the development of our applications. In the following, we briefly specify the diverse types of data, which we employ in this project.

First of all, we require data on the topology of the gas network. Therefore, we use a graph to model the network topology. Arcs of this graph represent single technical elements of the gas network, such as pipes, control valves, and compressor stations, while its nodes depict the junctions, sources, and sinks. Each of these network objects has a defined set of attributes describing the way they operate. Furthermore, there exist technical constraints like feasible pressure or velocity ranges in which the network elements have to be operated. Other characteristics include logical relations and dependencies between the individual elements. An example of such a specification has been given in Section 3 for the compressor stations.

To recommend future control measures for the single elements in the network, we need to know their current state. On the one hand, this involves the current physical network state in terms of pressure, flow, and other quantities, as well as the current mode of operation for active elements. On the other hand, we need to take already known future changes and limitations in the network into account, which may, for example, be caused by maintenance work. The resulting consequences for the network elements can range from tighter feasible operating ranges to a prescribed mode of operation for active elements.

For the gas demand forecasting system, we highly depend on the historical measurements of gas withdrawal and injection at the boundaries of the network. Moreover, we need data on external factors such as weather information as they may affect the gas demands in municipal areas. Likewise, for a subset of the boundary nodes, we have the nominations of the gas traders for the future. These are the announcements about the planned gas withdrawals from or injections into the network. However, they might not be accurate and are subject to change up to a few hours before their realization. To estimate the relationship between the predictors and predictand, being the final gas supply and demand, we need all the data mentioned above as well for the past.

4.2 The data interface

For the exchange of the described data, we designed an XML interface in close cooperation with our industry partner. In our previous project ForNe, we have already experienced the challenge of creating a gas network topology data set with above-presented diverse data types. With this in mind, we chose XML since it provides a formal and automated way to validate the input data for structural integrity using a corresponding schema XSD file, which is a preprocessing approach of category S2 described above. Hence, it enables us to detect data errors of that kind easily. Furthermore, having already validated input data significantly increased the overall robustness of the subsequent data parsing routines. The interface is based on the stationary GasLib format [6] and was adjusted to fit our needs and support the required information for the transient case. We plan to publish the xml interface and corresponding XSD-files in the next phase of the research campus MODAL.

4.3 Challenges we encountered

As shown in Section 4.1, we are interested in a diverse set of data. From our industry partner's point of view, the data is spread over different source systems since they are used in a particular context only. For example, general technical element properties

like the diameter of a pipe can be found in the system running the company's gas simulation software. At the same time, the nomination data is part of the system dealing contracts with gas traders. Each of these systems can have a different network representation, object identification system, and time granularity. Therefore, even if all the required data exists, it is most often not readily available.

In our case, one major challenge was the synchronization of these source systems, especially since some of them had never been used together. Given this setting, mapping objects from one system to another is not a trivial task, may require serious work, and might not even be possible in some cases. One example of this is the mapping between objects in the contractual trading system and the entry and exit points in the network. Here, some of the contractual nodes represent multiple nodes in the network. Hence, even if we have a guaranteed inflow at some contractual node for some time in the future, the inflow at the network level is not known in advance, since the trader can freely spread it over multiple nodes. Furthermore, even existing mappings might be non-trivial and hence cumbersome to deal with. In the source data of our industry partner, the simulation system has a very precise and detailed representation of the gas network. However, it contains no knowledge about the dependencies of the modes and configurations of the single active elements of the network, since these settings are part of the systems input data. These dependencies are stored in the network operators' system, which is based on an aggregated version of the network. Here, the topology is represented by fewer elements on a higher abstraction level to allow for a clear view of the primary transport connections. The knowledge to transform the dispatcher's commands into the modes and configurations on the level of single network elements is saved in a complex rule-based scripting system. Therefore, the set of all feasible combinations of different modes and configurations for the single elements is not explicitly present in the data, and its generation turns out to be quite complicated.

Connecting different source systems and comparing the corresponding data may lead to unexpected discrepancies. For example, we found that the operation points of single compressor machines are regularly outside of their feasible operating range. An example is given in Figure 1. From our project partner, we know that this is probably due to how the feasible operation range has been created. As described in Section 3 above, it was once fitted on different measurements of the running machine. Therefore, it accounts for neither the operation points attained while starting the compressor machine nor the changes of the machine's properties due to wear.

There was a second problem regarding the compressor related data. Some of the compressor machines' drives are powered by gas drawn directly from the network. For these drives, the consumption of gas is not measured - they just take what they need. Therefore, the energy consumption of these drives is simply not known.

Another major challenge results from the differences in the source systems regarding the time granularity mentioned above. These differences might lead to an initial network state being composed of mode and configuration values for the active elements from one point in time and pressure and flow values from another point in time, e.g., 3 minutes later. If one element switches its mode during these 3 minutes, we have a conflict in the initial state of this element. These inconsistencies are based on the general setup of the source systems and can hence not be fixed.

Although the use of XML schemata provides various benefits, it also involves some obstacles. First, there is the process of creating the initial version of the interface. Since the structure of the data has to be defined before the actual delivery, a considerable amount of work for the first analysis is needed here. Apart from identifying the necessary as well as potentially useful data types, we also have to take the general availability of the data and the corresponding effort of providing it into account.

However, even carefully created XML interfaces might need updates during the process of the project. Especially for research projects, such updates are usually unavoidable, considering that the demand for data might not be clear from the beginning,

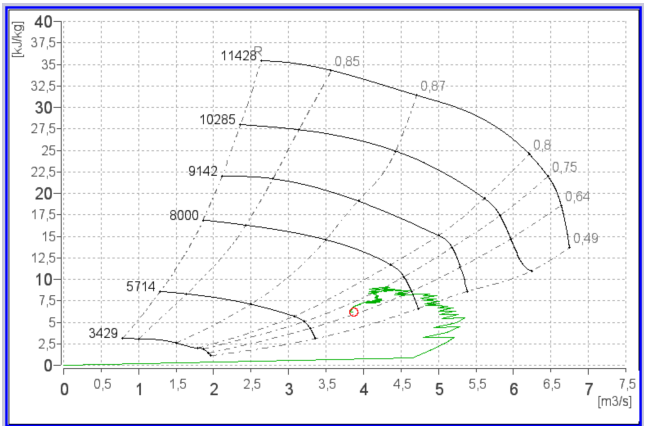

Figure 1: Example of erroneous input data. The grid in black displays the feasible operating range of a compressor unit. The red circle represents the current point of operation while the green line consists of all previously attained operation points. The majority of these previous operation points have been outside of the feasible operating range.

or novel approaches are developed that might require new types of data. For example, at the beginning of our project, the initial state's gas temperature was specified as an average value over the whole network. Later we discovered that this is too much of an approximation and changed this attribute to be given per node. Furthermore, we needed to add a new element type to the general topology description since the models of the standardized gas network element types could not capture the behavior of some metering units. Moreover, also smaller changes have been necessary, in cases of updated project requirements or data structures in the source systems, or when the initial schema definition was incomplete or erroneous. Finally, some of the originally specified data could not be delivered, since the effort of integration was beyond the scope of the project or the initial assumption about the availability of the corresponding data was wrong. All of these reasons contributed to the roughly two hundred adjustments, which we made between the initial and our current interface version. To illustrate the development of our interface definition and provide a rough measure for the work involved in the process, we present a general overview of the version history in Table 1.

Table 1: Version history of the XML interface in the GasLab 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

4.4 Solutions we found

Despite all of the above-mentioned challenges, the choice of the XML format for our data interface paid off in the end. A lot of data error can be easily detected in a standardized, automated fashion and do not have to be considered when parsing the

data. Therefore, we made the schema as specific as possible, for example, by fixing date formats and establishing value ranges like for geographical coordinates. All the corresponding checks are part of the preprocessing category S2.

To keep the number of changes as small as possible, good communication between all the different parties using the interface is essential. These range from specialists on the source system and transformation requirements, over researchers, to subject matter experts on gas dispatching. Furthermore, we allowed a partial deviation from the strict XML format for those file types, which are currently in development, manually created, or likely to change often for other reasons. For these files, we used the YAML format, which uses minimal syntax and is more human readable compared to XML. This gave us much flexibility and prevented a lot of interface changes, which was a more valuable contribution in these cases than the possibility to validate the data against a schema. We also integrated a free-text area into the XML for data that is only relevant for ourselves, e.g., to store debugging output or information from previous executions.

In case of a necessary change, we found that automated processes for the adjustment and release process reduce the necessary work considerably and decrease the likelihood of errors. Furthermore, we created tools to convert data from outdated interface formats to the newer ones. Thus, the data is easily accessible, and code for parsing does not need to cover all the different variants.

Independent of the structural validation provided by the XML schemata, we created a data checking tool for detecting potential logical errors in a systematic and reproducible way. Here, we evaluate the validity of different related values, which happens after the actual reading process. Hence, the tool is independent of the XML format. An example of such an error would be a pipe with a smaller length than the height difference of its end nodes. These kinds of errors are likely to happen, either by errors in the source systems, miscommunication, or bugs in the data transformation code. Without these checks, the errors might lead to failure or unexpected behavior in the subsequent parts of the code, which are often hard to detect and relate to the original error.

For each defined error, we specified either an automated way to handle the error appropriately, which is a preprocessing approach of category S2, or abort with a corresponding error message describing the source data's inconsistency. By communicating the error to our project partner as explained in the preprocessing category S3, they were able to find the problems' causes. However, sometimes correct data was just not available. This was the case for the above-described inconsistencies between the feasible operating range and the observed points of operation of some compressor machines. Since it is not possible to regularly repeat the measurements of each compressor machine in the network, gas network experts of our project partner finally created hand-tailored feasible operating ranges for each of the compressor machines. Similarly, we got estimates of the energy consumption of the gas-powered drives. These kinds of actions are part of the S4 category of preprocessing methods.

In Figure 2, we present statistics in terms of the number of warnings and errors found by the data checker in the regularly delivered input data over multiple months. Most interestingly, we changed the interface during this time, making it more uniform. As a result, the number of different issue types decreased by roughly two thirds. However, due to the new structure we were able to check more thoroughly and the absolute number of data errors increased significantly.

Another important lesson we learned from this project is that independent of the amount of effort spent in creating a consistent real-world data set, there might be some issues that will not be fixed. Either these problems are structural, or the cost to adjust the corresponding source systems or transformation is unreasonably high compared to the benefit of a correction. While the data can sometimes be manually adjusted, as in the case of compressor machines' feasible operating ranges, this is only possible if the errors do not regularly occur, and the number of single issues is manageable by hand.

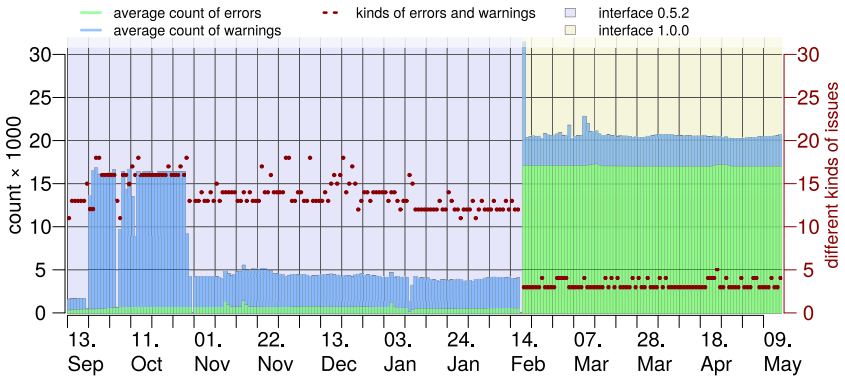

Figure 2: Data errors and warnings found by the data checker in the regularly delivered input data over time. The bars represent the total number of warnings and errors, while the red dots indicate the number of different types of issues(warnings and errors combined). The used interface version can be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

If this is not the case, the tools using the data have to find ways to deal with missing or erroneous data and should be as robust as possible regarding the potential lack of consistency. In general, for each error that turns out to be difficult to fix, enhancing the subsequent algorithms to not rely on the correctness of this type of data should always be considered as an option. As an example from our project, we had well understood statistical methods such as missing data processing or outlier detection at our disposal for the historical data used in the gas flow forecasting module. These methods are part of the preprocessing category S1. However, also the consistency of the initial state description could not be guaranteed, as explained in Section 4.3. We solved this issue by just ignoring the initial state problems, using the values as if they were correct, and enforcing the model constraints only for our future recommendations. However, in case of significant initial state errors, this approach might lead to an increased number of recommendations for the first time steps resulting from the necessity to switch to a consistent network state. A high number of these recommendations should be avoided since they just reflect technical problems of the tool and do not indicate actually needed control changes when comparing against the dispatchers' control decisions for the same network situation. Hence, careful monitoring of the results regarding this problem is needed. This approach can be interpreted as preprocessing in term of category S4 since we basically used the potentially inconsistent initial state values as a heuristic solution for the desired consistent initial state.

Using the measures we described here, we were able to tackle the challenges we described in Section 4.3 and used the data successfully in our decision support system as well as for research projects.

5 Example 2: Public data on Gas Networks

In the Horizon 2020 framework during the last three years [27], the EU invested more than 30 million Euros into research and innovation projects, some of them focusing on the transition towards a more sustainable energy pathway. In such a research environment, it is a common expectation that the gas transport network data is read-

ily available at institutions like ENTSO-G or national counterparts, for example, the *Vereinigung der Fernleitungsnetzbetreiber Gas e. V.* (FNB) in Germany. However, these organizations collect only data from the Transmission System Operators (TSOs) that are relevant for their analysis. They focus on investment decisions and the security of supply of energy and prepare ten-year network development plans [39, 40]. Consequently, they utilize high-level economic models that do not take into account the physics of gas in the transport network. As a result, readily available data sets of the gas transport network are not present in these organizations.

TSOs are obliged to publicly publish some information on their gas transport networks by the European Union and its national implementation on the transparency of gas transport networks. This information includes capacities and pressure bounds of nominated entry and exit nodes, the cumulative length of pipelines by diameter groups, and an illustration of networks on maps. For publishing this data, some TSOs utilize their transparency systems while others use documentation such as reports and web sites, or both. Because of the lack of a standard specification for publishing the transparency data, TSOs use different formats, naming conventions, and terms and conditions to publish this required information. Collecting relevant data from TSO data resources requires a substantial amount of effort, given that in Germany alone, there are already 16 TSOs by the time we write this paper.

We study the integration of the gas transport network in the multi-energy system, especially in terms of its connection to the electricity grid. During the transition towards being fossil-free, the energy system incorporates an increasing share of highly variable renewable energy sources. In times or areas where solar and wind power are less accessible, gas-powered-plants can produce energy with a short set-up time, hereafter referred to as G2P. This is in contrast to coal-powered-plants or nuclear power plants with long ramp-up and shut-down periods. Moreover, the gas transport network can serve as a flexibility via smart scheduling of compressor stations and the use of linepack. Likewise, P2G technologies convert excess electricity to hydrogen or ammonia, which, to a certain extent, can be induced to the gas transport network [41, 42, 43]. We aim to analyze the feasibility of decisions to build future energy systems, including mobility and heating systems, given the existing gas transport network infrastructure. The questions we seek to answer include, but are not limited to,

- whether it is feasible to transport/store the resulting gas, including adjustments from P2G and G2P, in the existing gas transport network, and storage facilities connected to the gas network until it is required,
- to what extent the gas transport network can be used as a flexibility to the energy system. For instance, the excess available electricity can be used to compress the gas in the network via electricity-driven compressor stations and store gas in the transport network itself or into storage facilities connected to the network.

Here, from our experience in the Horizon 2020 project plan4res, we will discuss what data is needed to study long-term planning of gas transport networks on a European scale and our challenges in obtaining these data. Finally, we present our strategies, which are mainly of preprocessing category type S4 and also S1, to cope with the lack of a comprehensive data set for this purpose.

5.1 What we needed

Although we deal with a long-term strategic planning problem, using an economic model or a minimum cost network flow model that is not considering gas pressure but only flow in the network is not an option. Because of the key characteristics of gas transport (see Section 3), such a model would provide only rough solutions that

are not sufficient to answer the questions we are interested in. Therefore, we adapt a physical-flow based stationary gas network optimization model [37, 44, 45] to our problem.

The required data for such a model as well as a corresponding data format is defined in the GasLib [6, 7]. We use the data format and attributes as defined by network description, compressor station description, and nomination data types of Gaslib.

- The network description data comprises the topology of the network and the technical data of all network elements such as pipes, control valves, valves, and compressor stations.
- The compressor station description data includes the complete and detailed description of all compressor stations that have been listed in the corresponding network description data, as also detailed in Section 3.
- The nomination data defines stationary nominations, i.e., balanced inflow and outflow scenarios for the entry and exit nodes.

We can categorize the network topology and compressor station data into mandatory data, whose lack of presence makes the model results unusable, and optional data, on which we can make assumptions using the information in public resources or literature, and summarized as follows.

- Mandatory data
  - geographical location of nodes and arcs,
  - location of compressor stations, control valves and valves,
  - diameter of pipelines,
  - locations of entry and exit nodes
  - types of facilities at the entries and exit nodes, i.e., storages, cross border points, consumers, etc.
- Optional data
  - gas composition at the entry nodes,
  - further technical details about pipelines such as roughness and heat transfer capacity; compressor stations such as configuration of compressors and drives, maximal compression ratios; control valve such as maximal pressure difference, and valves,
  - pressure bounds at nodes, especially at the entry and exit nodes.

The nomination data comprises the amount of gas flowing out from (into) the network at exit (entry) nodes. Since we are using a stationary gas network optimization model, the nominations must correspond to the stationary set-up: the total amount of gas inflow and outflow should be equal. Regarding our long-term planning problem, we need the nomination data of the entire region of interest, being Europe in our case. Since none of the network topology data is available on a European level and compiling them beyond the scope of the project, we reduced the area of interest to Germany. Once these data become available for other regions or the entire European scale, the concepts and models are ready to be applied on these scales. Therefore, we initially require the data for the amount of gas flowing into Germany via entry nodes and flowing out of Germany via exit nodes. The entry nodes include cross border points, where Germany imports gas, and indigenous production sites. The exit points consist of cross border points, where Germany exports gas, and final consumer points. Besides, we need the data of storage facilities, which serve as both entry and exit points to the gas transport network.

As observed from the list, compressor station data as detailed in Section 3 are optional. However, we have to locate the compressor stations in the gas transport network correctly.

5.2 Challenges we encountered

Our main challenge emerges from the lack of readily available gas transport network data in public resources. Therefore, we have to obtain the required data presented in Section 5.1 from various publicly available data sources to build a reliable and consistent data set precise enough to be used in our analysis. For this reason, we started working on collecting data from different data sources in addition to data provided by ENTSO-G and FNB via their web sites, reports, and data repositories [2, 46, 47, 48]. TSO web sites and data repositories, which are called transparency platforms (TP), constitute our main source of data other than those organizational data sources. In addition, we used the Aggregated Gas Storage Inventory Transparency Platform (AGSI+) of Gas Infrastructure Europe (GIE) [49, 50]. GIE is an organization of gas infrastructure operator companies across Europe that operate transmission pipelines, storage facilities, and LNG terminals.

The used publicly available gas transport network data are summarized in Table 2 by their source and format. The rows of the table present the required data, while the columns show the data sources grouped by type. Data in these sources are available in various formats: (i) a document from the web site, (ii) downloadable documents from TP, (iii) a format specific to TSO, (iv) visualization, (v) a table in a document. Additionally, the granularity of data varies: (i) per balancing zone, (ii) per gas system, (iii) per network cluster (NUTS3 Region), (iv) per interconnection point, (v) per important node, (vi) per pipeline diameter class, (vii) per pipeline, (viii) per storage, (ix) per storage per reserve amount. In the table, each available data is represented by its format type and granularity type, i.e., <Format type>/<Granularity type>.

Table 2: Publicly Available High-Pressure Gas Transport Network Data

Required Data | Data Source		Maps / Stylish Images		Transparency platform				TYNDP Scenarios		Web site
		ENTSO-G	TSOs	ENTSO-G	FNB	GIE	TSOs	ENTSO-G	FNB	TSOs
Gas properties	Gas quality type			DfTP/IP	DfTP/IP		TSOd/IN			
	GCV			DfTP/IP			TSOd/IN			
Infrastructure	Flow direction			DfTP/IP	DfTP/IP		TSOd/IN			
	Node facility types			DfTP/IP	DfTP/IP		TSOd/IN			
	Operator			DfTP/IP	DfTP/IP		TSOd/IN			
	Pressure bounds						TSOd/IN			
	Technical capacity of nodes			DfTP/IP	DfTP/IP		TSOd/IN			
Network Topology	Node locations	Vis/IP	Vis/IN							
	Pipeline diameter	Vis/P	Vis/P; Vis/-							Text/DN
	Pipelines (start-end nodes)	Vis/P	Vis/P							
	Pipelines (Length)									Text/DN
Nomination	Demand forecast							Table/BZ	Vis/NC	
	Supply forecast							Table/BZ	Vis/NC	
	Gas inflow			DfTP/IP			TSOd/IN			
	Gas outflow			DfTP/IP			TSOd/IN			
Storage specific	Insertion capacity					DfTP/Str				
	Insertion rate					DfTP/StrR				
	Present reserve					DfTP/Str				
	Withdraw capacity					DfTP/Str				
	Withdraw rate					DfTP/StrR				

Legend

<Format type> / <Granularity type>

Available format type: Text: Document from Web site; DfTP: Downloadable from TP; TSOd: TSO dependent format; Vis: Visualization; Table: Table in a document

Available granularity types: BZ: per balancing zone; GS: per gas system; NC: per network cluster (NUTS3 Region); IP: per interconnection point; IN: per important node; DN: per pipeline diameter class; P: per pipeline; Str: per storage; StrR: per storage per reserve amount

We can observe from Table 2 that the data sources are not mutually exclusive, i.e., there are entities shared by different data sources. Since the data sources use

various formats, shared entities are not necessarily identically named or labeled. Thus, consolidation of data from different sources into a single data set is needed.

To perform this consolidation, we have to understand the format and naming conventions used by the different data sources. The content and naming conventions of data in public sources depend on the purpose of the organization or the company owning the data. The naming conventions do not necessarily match each other or those we use in gas network optimization models. One of the differences results from naming different entities in different types of sources as *points*. Points in the TSO data context are the exit and the entry nodes of the TSO gas transport networks. This convention is intuitive in the gas network optimization context as well. However, points in ENTSO-G TP are the links connecting different types of facilities, TSO networks, and balancing zones. ENTSO-G uses a multi-layered approach for modeling the European gas transport network [51]. They only model the upper-most layer exit-entry network using a minimum cost network flow problem to route the supplied gas to meet the demand. Besides, they assume that any upper-most layer feasible flow entering/leaving the transport network can be transported through the physical gas transport network as long as it meets the capacity constraints provided by the TSOs and the balancing constraints. ENTSO-G employs a specific network topology [52] to model the upper-most layer as a directed graph. The nodes of this graph are components of the high-level gas network, such as demand-attached gas systems, storage facilities, LNG facilities, artificial supply nodes per country. The edges of this graph are unidirectional or bidirectional capacitated connections that correspond to paths in the real gas transport network.

Another confusion related to the naming convention results from the requirement to combine more than one TSO network. Not all of the entry or exit nodes of TSOs that provide a connection to other TSOs at market exchange areas or cross border points constitute an entry or exit point to a country-wide or EU-wide gas transport network. Some of them serve as transshipment nodes as they are intermediate to the country or EU-level network consisting of more than one TSO network. Hence, it is crucial to understand the purpose that TSO entry and exit nodes serve. For example, ENTSO-G provides cumulative demand data for each balancing zone that has to be dispatched to the relevant exit nodes of the TSOs in the respective balancing zone. Still, TSOs list the nodes that serve as an interconnection point to another TSO, which is not physically an exit point, as an exit point as well. Hence, the point capacities or point flows, as provided in ENTSO-G TP, cannot be directly matched to the amount of gas flowing into or out of the network via physical entry (exit) nodes as required in scenario files.

In addition to the difference in content, each TSO has its transparency systems or documentation. The data has to be obtained from those resources using various tools resulting in different formats, which is already a challenge for the German network. It obviously becomes a project in itself when compiling a European data set.

Another major challenge is the assignment of proper geographic location data to the network entities. TSO and organizational databases do not include any geographic location data of pipelines and nodes (see 2), except for information presented in pictures and maps that are not georeferenced. Additionally, pipeline diameters are only available from those visuals. Some inferences can be made on pipeline diameters using structural data on TSO web sites that incorporate cumulative length information of pipelines per diameter class. On the other side, Open Street Map (OSM) [53] provides some data for natural gas pipelines in Europe. Still, it is far from covering the entire grid with a considerable number of links missing, especially for Germany. Kunz et al. published a data set of the German gas transport network [54]. This data set includes the georeferences of a majority of pipelines in the German gas transport network. They used the pipeline maps and pictures provided by TSOs and organizations such as FNB, and data contained in public data sets to prepare this data. Besides,

Kunz et al. utilized several heuristic procedures using the structural pipeline data from the TSO web sites to estimate the unknown pipeline diameters [35] and modeled the German gas transport network as a capacitated minimum cost network flow problem. They computed the pipeline capacities using the diameter values, most of which they heuristically estimated, and assumption-based maximum pressure for pipeline diameter classes. Although working with the capacitated minimum cost network flow problem, we cannot find feasible solutions with this data set employing a stationary gas network optimization model with the nomination scenarios generated from the historical gas supply and demand data. During our analysis, we observed that the main problems of the data set causing this infeasibility are roughly estimated pipeline diameter data and lack of proper data on active components of the gas transport network such as compressor stations. Unfortunately, we have very sparse publicly available data regarding pipeline diameters, as mentioned above, and active components of the German gas transport network. For active components, we have a list of compressor stations with names, operators, and maximum power data [36] and an illustration of the location of these compressor stations in Germany. These can be complemented by information from the TSOs websites. However, the available data is very limited compared to our compressor data example in Section 3. For instance, for Germany, the available data is not more than the relative locations of compressor stations with respect to the gas pipes, the number of the compressor machines in a compressor station, and maximum power and drive type of those compressor machines [36]. Compared to the required data in Section 3, the feasible operating regions, compression efficiency, or possible configurations of the compressor machines in each compressor station are not available in the public data. The available compressor station data have to be associated with the network topology data using the names of the compressor station nodes. The rest of the required data must be modeled using appropriate assumptions and models by using an S4 type preprocessing approach. To study the potential of the gas transport network as a flexibility to the energy system, it is essential to properly estimate the network's existing capacity and make inferences on the excess/lack of capacity of the gas transport network.

To prepare adequate scenarios, further transformation, not only involving network topology but also the nominations, is necessary. For example, cumulative supply and demand forecasts are provided for balancing zones in ENTSO-G databases. The amount of gas entering and leaving the German gas transport network via entry and exit nodes should be computed using these forecasts. As a caveat, the supply and demand forecast is only provided on a yearly resolution. But, the temporal resolution for our studies regarding the integration of gas network to electricity network is a day or an hour.

Finally, the public data resources only contain pressure bounds for the main pipelines and important nodes of a limited number of TSOs. Remaining pressure bounds have to be consistently estimated using available data to allow for a more precise analysis.

5.3 Solutions/Remedies we found

Our remedies involve the consolidation of public data that is connected with sparse availability by using data preprocessing. We employ S1 type of preprocessing, such as deleting/fixing errors and replacing missing values by data set comparison or computations of simple gas dynamics. However, because of the geospatially or temporally non-homogeneous data provided in public resources, we heavily rely on S4 type preprocessing. We use heuristics or mathematical models that use educated assumptions to compile a consistent data set by exploiting the connections in different public data sets. Thus, we use the S4 type preprocessing approach in our solutions presented in this section unless it is stated explicitly as the method uses the S1 type preprocessing

approach.

As a first remedy, we have examined the data related documentation of organizations and companies that provide public data carefully, with their purpose of using and publishing the data in mind, to better and thoroughly understand the content of the data. During this study, it is essential to read and understand how and why the data is collected and published by a particular organization or company before making any use of their data, not only from available data documentation but also from their analysis reports and publications. Hence, knowledge of how gas markets and gas transport networks operate and expertise in gas network optimization models are required to draw correct conclusions from the available data. This study lays the foundations of assumptions that we use in our heuristics and models that we employ in our S4 type data preprocessing. As a result, we have encountered several data transformation requirements and have developed models and methods for collecting and transforming data. We have already applied some of the methods. However, it is ongoing work to improve the data quality to achieve even more reliable results.

We used the network topology data of the German gas transport network data by Kunz et al. as a basis for our efforts to build a German gas transport network [54, 35] and augmented this data with the capacity and pressure bounds of important nodes obtained from the TSO and FNB data repositories. We utilized semi-automated methods to collect data from those publicly available resources and joined those data using text- and feature-based SQL-queries and scripts comparing text-fields such as names or operators. Joining different data sets also helped us to associate ENTSO-G data with the data set contained in [54, 35] . We improved the augmented data set by estimating the maximum pressure values of pipelines utilizing the node pressure bounds, which is of S1 type preprocessing.

Ongoing works include improving data by adding compressor stations and fixing inaccuracies in pipeline diameters. Associating the geographic locations of the network topology with the listed information of [36] will replace rough assumptions on the characteristics of compressor stations, but requires manual work. Since only a limited amount of data of compressor stations is available, we make educated assumptions for the type, feasible operating range, and efficiency of compressor machines, and technically possible configurations of compressor machines in the compressor stations. These assumptions are mainly based on our knowledge of the detailed compressor station models that have been used in our short term operation problem studies. We verify the assumptions using our mathematical models to understand whether the available and assumed data together constitute a valid compressor station model. A network optimization-based methodology is employed to improve the estimates of pipeline diameters. TSO data on pipelines are given as diameter classes, i.e., if a pipeline is of a particular class, then its diameter belongs to a predetermined range. To estimate the diameters, we have to match the pipelines in the network topology data with these diameter classes. This matching should result in the pipeline capacities complying with interconnection point capacities in ENTSO-G and TSO TPs, while the total length of pipelines by diameter classes does not exceed the total length given by each TSO's structural pipeline data. This approach represents a multi-facility network design model.

In parallel, we are working on extracting pipeline data from non-georeferenced maps or stylish images provided by ENTSO-G and TSOs. We make use of some tailored image processing and image registration methodologies to get network data from these images. Our first results show that these methods reduce the error and require less amount of effort compared to readily available georeferencing tools provided by geographical information systems (GIS).

In addition to our studies to improve the network topology data and extract pipeline data from open sources, eliciting scenarios for nominations using realistic data is necessary for the consistency of the analysis. Here, we focus on dispatching

yearly supply and demand forecasts provided by ENTSO-G at the country or balancing zone level. The forecast data should be spatially and temporally disaggregated. Thus, we developed a three-step approach, which is also presented in Figure 3.

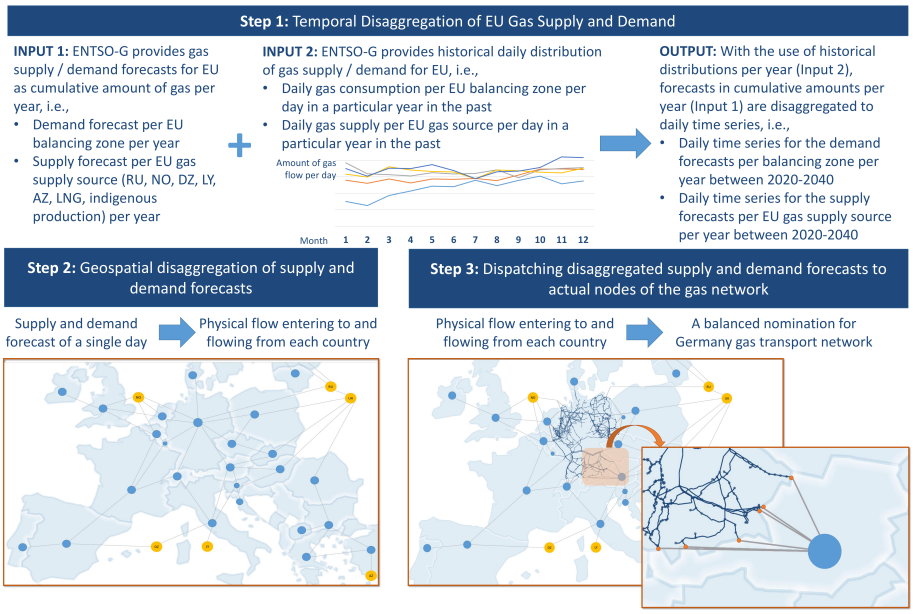

Figure 3: A Three-Step Procedure for Generating Nominations

i. Temporal disaggregation of yearly supply and demand forecasts: Yearly supply and demand forecasts are available by balancing zone on the European-level [39]. Additionally, historical data of supply and demand is available on a daily and hourly basis [46]. We require daily and hourly forecasts. We use an S1 type preprocessing approach and employ a time-series transformation to disaggregate the yearly forecasts of a single balancing zone to days of the year, based on a distribution obtained by the historical data.

ii. Geospatial disaggregation of supply and demand forecasts: ENTSO-G provides European level supply and demand forecasts. We re-modeled ENTSO-G's network topology used by their Network Modelling Tool (NeMo), whose details are provided in Section 5.2, to dispatch the European level supply and demand forecasts provided by ENTSO-G to the balancing zones and interconnection points of Germany. Our model is a capacitated minimum cost network flow model, but it has additional constraints for balanced gas inflow and outflow of balancing zones and countries. It differs from ENTSO-G's model since ENTSO-G uses the model to find minimum cost expansion plan on capacities, while our model identifies a routing for EU gas supply that minimizes total demand curtailment for all European Union countries. Given the forecasts by country as well as interconnection capacities, the model finds the amount of gas entering and leaving the German gas transport network via particular interconnection points to other countries, storages, indigenous production zones, and final consumers. Since the model accounts for demand curtailment needed by countries in case of a shortage situation, this model can also create a balanced scenario as required by the stationary gas network optimization model.

iii. Dispatching disaggregated supply and demand forecasts to actual nodes of the gas transport network: The supply and demand forecasts are spatially disaggregated to Germany's interconnection points in the previous step. However, interconnection points cannot be associated with physical gas transport network nodes on a one-to-one basis. We use a capacitated network flow model to dispatch the supply and demand forecasts of interconnection points to the nodes of the German gas transport network to generate a balanced nomination scenario. The nodes and arcs of the network model represent actual nodes and pipelines of the physical gas transport network. The capacities of pipelines are assigned as capacities of bidirectional arcs between nodes. The interconnection points are modeled by artificial nodes and linked to the associated entry or exit nodes of the gas transport network via unidirectional and uncapacitated arcs. The association of interconnection points and physical network nodes, which is either determined automatically when the data from different sources are joined or manually, is given in the data set. We are currently investigating methods to automate the manual associations by making use of the network visuals.

The network topology and compressor stations data are fixed for a given gas transport network unless there is an update in the network topology, i.e., network expansion by adding new pipelines or adding/upgrading components such as compressor stations. Although our studies are based on the current network topology, the network topology expansion data can be incorporated to our data set with further efforts. Ten-year network development plans prepared by ENTSO-G for European Union or by FNB for Germany can be used as a basis for this improvement.

6 Conclusion

For Example 1, we have integrated and exchanged data that describes gas transport scenarios. With the support of our partners from industry, resources are exchanged and recorded every 3 to 30 minutes. This data covers the entire history of over seven years of one of the largest German gas networks, containing approximately 12,000 km of pipelines. On this data, we proceed with our forecasting and optimization algorithms to achieve many valuable research goals, see, e.g., [24, 25, 26].

In Example 2, we have built a gas network data-set for Germany from public data sources by using data scattered around in different formats, including pictures, lists, and specification sheets. Besides, we have integrated this data with the supply and demand forecasts from ENTSO-G, enabling our research on gas transport networks to add flexibility to European energy systems optimization. Once we complete our studies, we will publish an open access data set prepared in the context of plan4res.

Publishing Open Access data is only possible when it is also allowed by the data owner. For instance, data can be confidential as in Example 1, or data sources have terms and conditions for data utilization preventing third parties from sharing the data even after processing it, as in the case of [17]. However, comparative research needs common problem instances and related data-sets. Thus, in Example 2, we are paying the utmost attention to prepare a data set that we can share publicly.

In total, over both case studies, we have spent a considerable number of person-years preparing consistent data sets. It took much effort regardless of whether using company data or public data. Additionally, domain knowledge regarding gas network optimization models, as well as the operation of gas markets and gas networks, was necessary. Open accessible, well-prepared, and realistic data sets can be very valuable. 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 [55]. Such benchmarks are essential to compare different modeling approaches and different scenarios. MIPLIB [56] is an excellent example of how such benchmark libraries boost research

and collaboration in a specific research area. For example, [17] has been cited more than 100 times, although not all of these papers focus on gas networks. The [35] data set has more than 1000 views and more than 250 downloads. Hence, we give utmost importance to complement the GasLib gas network instances [6, 7], which have been cited by more than 50 papers in the last five years, with new realistic gas transport network instances such as presented in 5. We believe that these efforts on the benchmark library will synergize both the gas network optimization and energy systems optimization research.

In both case studies, we would have profited from information available to individual companies or (non-)governmental organizations. Unfavorably, these data are not shared with the public. Furthermore, we were restricted in our possibilities to share the preprocessed data-sets we prepared. However, until there is a change in Open Access regulations and companies' attitudes towards data sharing, the lessons learned here will guide research groups following us on this endeavor.

Finally, the envisioned transition towards more renewable energy sources may involve mixing other gases such as hydrogen or biogas into the natural gas transport network. These changes will lead to further challenges to the gas network operators, which require innovative algorithmic solutions, based on even more data.

Acknowledgments

The work for the results reported in Section 4 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 5 are part of a project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 773897.

Abbreviations

ForNe	Forschungskooperation Netzoptimierung (Research Cooperation Network Optimization)
ENTSO-G	European Network of Transmission System Operators for Gas
FNB	Vereinigung der Fernleitungsnetzbetreiber Gas e.V.
G2P	Gas-to-Power
P2G	Power-to-Gas
RES	Renewable Energy Sources
TP	Transparency Platform
TSO	Transmission System Operator

References

[1] Regulation (EC) No 715/2009 of the European Parliament and of the Council on conditions for access to the natural gas transmission networks. Available online: `https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:211:0036:0054:en:PDF` (accessed on 19.05.2020).

[2] ENTSO-G Website. Available online: `https://www.entsog.eu/` (accessed on 19.03.2020).

[3] ENTSO-G Transmission Capacity Map. Available online: `https://www.entsog.eu/sites/default/files/2020-01/ENTSOG_CAP_2019_A0_1189x841_FULL_401.pdf` (accessed on 22.03.2020).

[4] ENTSO-G/GIE System Development Map. Available online: https://www.entsog.eu/sites/default/files/2020-01/ENTSOG_GIE_SYSDEV_2018-2019_1600x1200_FULL_063_clean.pdf (accessed on 22.03.2020).

[5] ForNe Project Website. Available online: https://www.zib.de/projects/forne-research-cooperation-network-optimization (accessed on 22.03.2020)

[6] Schmidt, M.; Aßmann, D.; Burlacu, R.; Humpola, J.; Joormann, I.; Kanelakis, N.; Koch, T.; Oucherif, D.; Pfetsch, M.E.; Schewe, L.; Schwarz, R.; Sirvent, M. GasLib—A Library of Gas Network Instances. *Data*, **2017**, *2*, 40.

[7] GasLib - a library of gas network instances. (2018). Available online: https://gaslib.zib.de (accessed on 16.04.2020).

[8] BDEW-Energiemarkt Deutschland 2020. Available online: https://www.bdew.de/service/publikationen/bdew-energiemarkt-deutschland-2020 (accessed on 19.06.2020).

[9] OGE – Company – Our history. Available online: https://oge.net/en/us/company/our-history (accessed on 19.06.2020).

[10] LKD-EU Project Website. Available online: https://www.diw.de/de/diw_01.c.537097.de/projekte/langfristige_planung_und_kurzfristige_optimierung_des_elektrizitaetssystems_in_deutschland_im_europaeischen_kontext_lkd_eu.html (accessed on 22.03.2020).

[11] SciGRID Gas Project Website. Available online: https://www.gas.scigrid.de (accessed on 19.03.2020).

[12] Orlowski, S.; Wessäly, R.; Pióro, M.; Tomaszewski, A. SNDlib 1.0—Survivable Network Design Library. *Networks*, **2010**, *55(3)*, 276-286.

[13] Arnolda, F.; Gendreau, M.; Sörensen, K. Efficiently solving very large-scale routing problems. *Computers & Operations Research*, **2019**, *107*,32-42.

[14] Babaeinejadsarookolaee, S.; Birchfield, A.; Christie, R.D.; Coffrin, c.; DeMarco, C.; Diao, R.; Michael Ferris, M.; Fliscounakis, S.; Greene, S.; Huang, R.; Josz, C.; Korab, R.; Lesieutre, B.; Maeght, J.; Molzahn, D.K.; Overbye, T. J.; Panciatici, P.; Park, B.; Snodgrass, J.; Zimmerman, R. The Power Grid Library for Benchmarking AC Optimal Power Flow Algorithms.**2019**, *arXiv preprint arXiv:1908.02788*.

[15] Kotsiantis, S. B.; Kanellopoulos, D.; Pintelas, P. E. Data preprocessing for supervised leaning. *International Journal of Computer Science*, **2006**, *1(2)*, 111-117.

[16] Otey, M.E.; Ghoting, A.; Parthasarathy, S. Fast Distributed Outlier Detection in Mixed-Attribute Data Sets. *Data Mining and Knowledge Discovery*, **2006**, *12*, 203–228.

[17] Carvalho, R.; Buzna, L.; Bono, F.; Gutiérrez, E. Robustness of Trans-European Gas Networks. *Physical review E*, **2009**, *80*, 016106.

[18] Karpievitch, Y.V.; Dabney, A.R.; Smith, R.D. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics*, **2012**, *13*, S5.

[19] Rettig, L.; Mourad, K.;Cudré-Mauroux, P.; Piorkówski, M. Online Anomaly Detection over Big Data Streams. In *Applied Data Science*; Braschler, M.; Stadelmann, T.; Stockinger, K.; Eds.; Springer, Cham, 2019.

[20] Benner, P.; Grundel, S.; Himpe, C.; Huck, C.; Streubel, T.; Tischendorf, C. Gas Network Benchmark Models. In *Applications of Differential-Algebraic Equations: Examples and Benchmarks* ; Campbell, S.; Ilchmann, A.; Mehrmann, V.; Reis, T.; Eds.; Springer, Cham, 2018.

[21] Hiller, B.; Hayn, C.; Heitsch, H.; Henrion, R.; Leövey, H.; Möller, A.; Römisch, W.; Methods for verifying booked capacities. In *Evaluating Gas Network Capacities*; Koch, T.; Hiller, B.; Pfetsch, M.E.; Schewe, L.; Eds.; SIAM-MOS series on Optimization: Berlin, Germany, 2015; pp 291-315.

[22] Li, B.; Simulation and capacity calculation in real German and European interconnected gas transport systems, Cuvillier Verlag, Göttingen, 2012.

[23] Research Campus Modal WebSite. Available online: `http://forschungscampus-modal.de/en/about-us/gas-lab-en/` (accessed on 22.03.2020)

[24] Petkovic, M.; Chen, Y.; Gamrath, I.; Gotzes, U.; Hadjidimitriou, N.S.; Zittel, J.; Koch, T. A Hybrid Approach for High Precision Prediction of Gas Flows. *ZIB-Report 19-26* **2019**.

[25] Hoppmann, K.; Hennings, F.; Lenz, R.; Gotzes, U.; Heinecke, N.; Spreckelsen, K.; Koch, T. Optimal Operation of Transient Gas Transport Networks. *ZIB-Report 19-23* **2019**.

[26] Hennings, F.; Anderson, L.; Hoppmann, K.; Turner, M.; Koch, T. Controlling transient gas flow in real-world pipeline intersection areas. *ZIB-Report 19-24* **2019**.

[27] European Commission Horizon 2020 Web Site. Available online: `https://ec.europa.eu/programmes/horizon2020/en` (accessed on 08.04.2020).

[28] plan4res Project Website. Available online: `https://www.plan4res.eu` (accessed on 19.03.2020).

[29] Most, D.; Giannelos, S.; Yueksel-Erguen, I.; Beulertz, D.; Haus, U.; Charousset-Brignol, S.; Frangioni, A. A Novel Modular Optimization Framework for Modelling Investment and Operation of Energy Systems at European Level. *ZIB-Report 20-08* **2020**.

[30] Qi, C. Big data management in the mining industry. *International Journal of Minerals, Metallurgy and Materials*, **2020**, *27(2)*, 131-139.

[31] Mohammadpoor, M.; Farshid, T. Big Data analytics in oil and gas industry: An emerging trend. *Petroleum*, **2018**, `https://doi.org/10.1016/j.petlm.2018.11.001`.

[32] Kim, T.; Li, W.; Behm, A.; Cetindil, I.; Vernica, R.; Borkar, V.; Carey, M.J.; Li, C. Similarity query support in big data management systems. *Information Systems*, **2020**, *88*, 101455.

[33] Wahab, O.A.; Cohen, R.; Bentahar, J.; Otrok, H.; Mourad, A.; Rjoub, G. An Endorsement-based Trust Bootstrapping Approach for Newcomer Cloud Services. *Information Sciences*, **2020**, *527*, 159-175.

[34] Rjoub, G.; Bentahar, J.; Wahab, O.A. BigTrustScheduling: Trust-aware big data task scheduling approach in cloud computing environments. *Future Generation Computer Systems*, **2019**, `https://doi.org/10.1016/j.future.2019.11.019`.

[35] Kunz, F.; Kendziorski, M.; Schill, W.P; Weibezahn, J.; Zepter, J.; von Hirschhausen, C.; Hauser, P.; Zech, M.; Moest, D.; Heidari, S.;Felten, B.; Weber, C. **2017**. Data Documentation 92: Electricity, Heat, and Gas Sector Data for Modeling the German System. Available online: `https://www.diw.de/documents/publikationen/73/diw_01.c.574130.de/diw_datadoc_2017-092.pdf` (accessed on 19.03.2020).

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

[37] Koch, T.; Hiller, B.; Pfetsch, M.E.; Schewe, L. Evaluating Gas Network Capacities, MOS-SIAM Series on Optimization, Berlin, 2015. https://doi.org/10.1137/1.9781611973693

[38] Walther, T.; Hiller, B. Modelling compressor stations in gas networks. *ZIB Report 17-67* **2017**.

[39] ENTSO-G Ten Year Network Development Plans. Available online: `https://www.entsog.eu/tyndp#` (accessed on 08.04.2020).

[40] FNB Netzentwicklungsplan 2018. Available online: `https://www.fnb-gas.de/netzentwicklungsplan/netzentwicklungsplaene/netzentwicklungsplan-2018/` (accessed on 08.04.2020).

[41] ENSO-G and ENTSO-E. Power to Gas-A Sector Coupling Perspective. Available online: `https://www.entsoe.eu/2018/10/15/power-to-gas-a-sector-coupling-perspective/` (accessed on 08.04.2020).

[42] Blanco, H.; Faaij, A. 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*; **2018**, *81*, 1049–1086.

[43] Schiebahn, S.; Grube, T.; Robinius, M.; Tietze, V.; Kumar, B.; Stolten, D. Power to gas: Technological overview, systems analysis and economic assessment for a case study in Germany.*International Journal of Hydrogen Energy*; **2015**, *40(12)*, 4285–4294.

[44] Fuegenschuh, A.; Geissler, B.; Gollmer, R.; Morsi, A.; Pfetsch, M. E.; Roevekamp, J.; Schmidt, M.; Spreckelsen, K.; Steinbach, M. C. Gas network elements. In *Evaluating Gas Network Capacities*; Koch, T.; Hiller, B.; Pfetsch, M.E.; Schewe, L.; Eds.; SIAM-MOS series on Optimization: Berlin, Germany, 2015; pp 17–44.

[45] Pfetsch, M.E.; Fuegenschuh, A.; Geissler, B.; Geissler, N.,Gollmer, R.; Hiller, B.; Humpola, J.; Koch, T.; Lehmann, T.,Martin, A.; Morsi, A.,Rovekamp, J.; Schewe, L.; Schmidt, M.; Schultz, R.; Schwarz, R.; Schweiger, J.; Stangl, C.; Steinbach, M. C.; Vigerske, S.; Willert, B. M. Validation of nominations in gas network optimization: models, methods, and solutions, *Optimization Methods and Software*, **2015**, *30*, 15-53.

[46] ENTSO-G Transparency Platform. Available online: `https://transparency.entsog.eu/` (accessed on 19.03.2020).

[47] FNB-Gas Website. Available online: `https://www.fnb-gas.de/` (accessed on 19.03.2020).

[48] FNB-Gas Network Development Plan Data Repository. Available online: `https://www.nep-gas-datenbank.de:8080/app/#\protect\leavevmode@ifvmode\kern-.1667em\relax/` (accessed on 19.03.2020).

[49] GIE Website. Available online: `https://www.gie.eu/` (accessed on 01.04.2020).

[50] AGSI+ Transparency Platform. Available online: `https://agsi.gie.eu/#/` (accessed on 01.04.2020).

[51] ENTSOG Union-wide Security of Supply Simulation Report. (2017). Available online: `https://www.entsog.eu/security-of-supply-simulation#union-wide-simulation-of-supply-and-infrastructure-disruption-scenarios` (accessed on 19.03.2020).

[52] Basic features of ENTSOG network and market tool. (2017). Available online: `https://www.entsog.eu/sites/default/files/2019-10/Basic%20features%20of%20ENTSOG%20TYNDP%20tool.pdf` (accessed on 19.03.2020).

[53] Open street Map. Available online: https://www.openstreetmap.org (accessed on 19.03.2020).

[54] Kunz, F.; Weibezahn, J.; Hauser, P.; Heidari, S.; Schill, W. P.; Felten, B.; Kendziorski, M.; Zech, M.; Zepter, J.; von Hirschhausen, C.;Moest, D.; Weber, Christoph. **2017**. Reference Data Set: Electricity, Heat, and Gas Sector Data for Modeling the German System (Version 1.0.0) [Data set]. Available online:https://doi.org/10.5281/zenodo.1044463 (accessed on 19.03.2020).

[55] National Institute of Standards and Technology. (2016). NIST Handprinted Forms and Characters Database [Data set]. Available online https://doi.org/10.18434/T4H01C (accessed on 22.05.2020).

[56] MIPLIB 2017 – The Mixed Integer Programming Library (2017). Available online: https://miplib.zib.de (accessed on 15.04.2020).