SciGRID_gas: The raw EMAP data set

Release 1.0

J.C. Diettrich & A. Pluta & W. Medjroubi

Jul 22, 2020

CONTENTS

How to cite

J.C. Diettrich, A. Pluta, W. Medrjoubi
SciGRID_gas: The raw EMAP data set
DLR Institute for Networked Energy Systems
Germany
doi: ???

Impressum

DLR Institute for Networked Energy Systems
Carl-von-Ossietzky-Str. 15
26129 Oldenburg
Germany
Tel.: +49 (441) 999 060

LIST OF FIGURES

LIST OF TABLES

Summary

The goal of SciGRID_gas is to develop methods to create an automated process that can generate a gas transmission network data set for Europe. Gas transmission networks are fundamental for simulations by the gas transmission modelling community, to derive major dynamic characteristics. Such simulations have a large scope of application, for example, they can be used to perform case scenarios, to model the gas consumption, to minimize leakages and to optimize overall gas distribution strategies. The focus of SciGRID_gas will be on the European transmission gas network, but the principal methods will also be applicable to other geographic regions.

Data required for such models are the gas facilities, such as compressor stations, LNG terminals, pipelines, etc. One needs to know their locations, in addition to a large range of attributes, such as pipeline diameter and capacity, compressor capacity, configuration, etc. Most of this data is not freely available. However throughout the SciGRID_gas project it was determined, that data can be found and grouped into two fundamental different groups: a) OSM data, and b) non-OSM data. The OSM data consists of geo-referenced facility data that is stored in the OpenStreetMap (OSM) data base, and is freely available. However, the OSM data set currently contains hardly any other information than the location of the facilities. The Non-OSM data set can fill some of those gaps, by supplying information such as pipeline diameter, compressor capacity and more. Part of the SciGRID_gas project is to mine and collate such data, and combine it with the OSM data set. In addition heuristic tools are required to fill data gaps, so that a complete gas network data set can be generated.

Here, this document describes one of the non-OSM data set, called the "EMap" data set, which was generated from an EntsoG PDF map [EntsoG20]. This document here explains the origin and structure of this single data sets, and how the data set was generated.

In this document, the chapter "Introduction" will supply some background information on the SciGRID_gas project, followed by the chapter "Data structure", that gives a detailed description of the data structure that is being used in the SciGRID_gas project. Chapter "Data sources" describes the EMap data set.

The appendix contains a glossary, references, location name alterations convention and finishes with the table of country abbreviation.

CHAPTER

ONE

INTRODUCTION

SciGRID_gas is a three-year project funded by the German Federal Ministry for Economic Affairs and Energy [BMWi20] within the funding of the 6. Energieforschungsprogramm der Bundesregierung [BMWi11].

The goal of SciGRID_gas is to develop methods to generate and provide an open-source gas network data set and code. Gas transmission network data sets are fundamental for the simulations of the gas transmission within a network. Such simulations have a large scope of application, for example, they can be used to preform case scenarios, to model the gas consumption, to detect leaks and to optimize overall gas distribution strategies. The focus of SciGRID_gas will be the generation of a data set for the European Gas Transmission Network, but the principal methods will also be applicable to other geographic regions.

Both the resulting method code and the derived data will be published free of charge under appropriate open-source licenses in the course of the project. This transparent data policy shall also help new potential actors in gas transmission modelling, which currently do not possess reliable data of the European Gas Transmission Network. It is further planned to create an interface to [MMK16] or heat transmission networks. Simulations on coupled networks are of major importance to the realization of the German *Energiewende*. They will help to understand mutual influences between energy networks, increase their general performance and minimize possible outages to name just a few applications.

This project was initiated, and is managed and conducted by DLR Institute for Networked Energy Systems.

1.1 Project information

- **Project title**: Open Source Reference Model of European Gas Transport Networks for Scientific Studies on Sector Coupling (*Offenes Referenzmodell europäischer Gastransportnetze für wissenschaftliche Untersuchungen zur Sektorkopplung*)
- **Acronym**: SciGRID_gas (Scientific GRID gas)
- **Funding period**: January 2018 - December 2020
- **Funding agency**: Federal Ministry for Economic Affairs and Energy (*Bundesministerium für Wirtschaft und Energie*), Germany
- **Funding code**: Funding Code: 03ET4063
- **Project partner**: DLR Institute for Networked Energy Systems

Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

1.2 Background

As of today, only limited data of the facilities of the European Gas Transmission Networks is publicly available, even for non-commercial research and related purposes. The lack of such data renders attempts to verify, compare and validate high resolution energy system models difficult, if not impossible. The main reason for such sparse gas facility data is often the unwillingness of transmission system operators (TSOs) to release such commercially sensitive data. Regulations by EU and other lawmakers are forcing the TSOs to release some data. However, such data is sparse, and too often not clearly understandable for non-commercial operators, such as scientists.

Hence, details of the gas transmission network facilities and their properties are currently only integrated in in-house gas transmission models which are not publicly available. Thus, assumptions, simplifications and the degree of abstraction involved in such models are unknown and often undocumented. However, for scientific research those data sets and assumptions are needed, and consequently the learning curve in the construction of public available network models is rather low. In addition, the commercially sensitivity also hampers any (scientific) discussion on the underlying modelling approaches, procedures and simulation optimization results. At the same time, the outputs of energy system models take an important role in the decision making process concerning future sustainable technologies and energy strategies. Recent examples of such strategies are the ones under debate and discussion for the Energiewende [BundesregierungDeutschland20] in Germany.

In this framework, the SciGRID_gas project initiated by the research centre DLR Institute of Networked Energy Systems in Oldenburg aims at building an open source model of the European Gas Transmission network. Releasing SciGRID_gas as open-source is an attempt to make reliable data on the gas transmission network available. Appropriate (open) licenses attached to gas transmission network data ensures that established models and their assumptions can be published, discussed and validated in a well-defined and self-consistent manner. In addition to the gas transmission network data, the Python software developed for building the model SciGRID_gas are published under the GPLv3 license.

The main purpose of the SciGRID_gas project is therefore to open the door to new gas transmission network models and innovative ideas in energy system modelling by providing freely available and well-documented data on the European gas transmission network.

The input data itself is based on data available from openstreetmap.org (OSM) under the Open Database License (ODbL) as well as Non-OSM data gathered from different sources, such as Wikipedia pages, fact sheets from TSOs or even newspaper articles.

The main workload of this project is to:

- retrieve the OSM and Non-OSM data sets for the gas infrastructure
- merge all available data sets

- build a gas transmission component data set
- generate missing data using heuristic methods
- remove all gas facilities, that are not connected to pipelines.

CHAPTER

TWO

DATA STRUCTURE

A well designed and documented data structure is fundamental in any large scale project. Good data structure in combination with tools, based on algorithms, improve the performance of any project output.

This structure needs to represent the gas flow facilities as good as possible, Hence, it needs to include components, such as pipelines, compressors, etc. A finite number of components have been identified, that are required as building blocks of a gas network. In addition each component will contain attributes, such as pipeline diameter, maximal operating pressure, maximal capacity, number of turbines etc.

It is anticipated, that the adopted data structure can be implemented in different types of gas flow models and will be used by the research community for topics, such as sector coupling or identifying gas transmission bottlenecks.

Within the SciGRID_gas project, the structure of the data model is part of classes defined within the Python code. Alterations may occur over the duration of the project, but it is envisaged, that those will be small, and that compatibility will be assured.

The goal of this section is to describe in details the data structure that has been adopted and implemented into the Python code. This will be important in understanding other aspects of this document, such as exporting the data into CSV files or generating missing values.

Prior to the description of the data structure, the overall pathway of the data flow within the SciGRID_gas project will be explained, as it is believed, that such overview will help the reader.

2.1 Data structure description

This section contains information about the SciGRID_gas data structure, the format, and the code that can be used to import publicly available data into the project, so that it can be used in subsequent steps. Paramount for an understanding of the data structure is a good understanding of the terminology used throughout this section and the document in general. Hence, terminology will be introduced in the following sub-section.

2.1.1 Terminology

Throughout this document certain terms will be used, which will be described below and summarized as a picture in Figure 2.1.

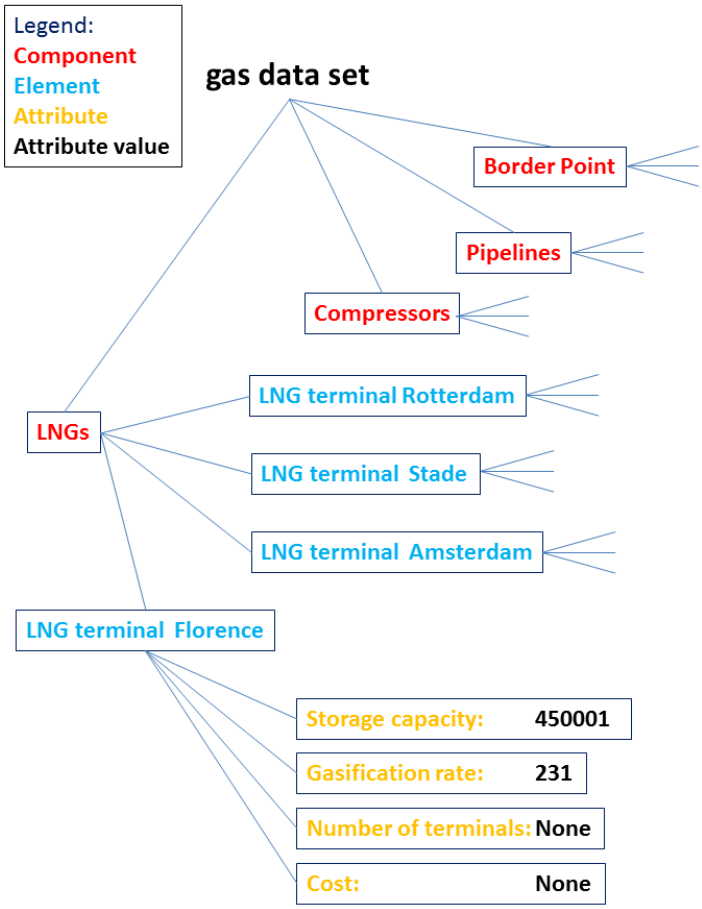

Figure 2.1: Data structure for the SciGRID_gas data set

Gas transmission network

The term "gas transmission network" describes the physical gas transmission grid. This does not include the distribution of gas through gas distribution companies, but includes the long distance transmission of gas from producer countries to consumer countries, as carried out by the Transmission System Operators (TSO) [Wik20g].

Gas component data set

The term "gas component data set" is used for all raw data of objects/facilities that have been loaded using SciGRID_gas tools into a Python environment. Gas component data sets are used as input into our SciGRID project. Several data sources can be loaded as gas component data sets, and then combined into a single gas component data set. However, not all elements (e.g. compressors) must be connected to pipelines, Hence, such a data set is referred to as a "gas component data set", and the emphasis is on the term **component**.

Gas network data set

A "gas component data set" can be converted into a "gas network data set", by connecting all non-pipeline elements to nodes and all nodes are connected to pipelines, and as part of the process all network islands have been connected or removed, resulting in a single network. Therefore the network contains nodes and edges which are coherently connected, and all objects with the exception of pipelines are associated with nodes in this network, whereas pipelines are associated with edges. Hence, the emphasis here is on the term **network**.

Component

There are several component types in a gas transmission network, such as compressors, LNG terminals, or pipelines. In Figure 2.1 they are coloured red. Hence, whenever the word "component" is mentioned, it refers to one of these components. There are roughly a dozen different components that will form a gas network data set. They will be briefly explained below.

Element

The term "element" refers to individual facilities, e.g. the LNG Terminal in Rotterdam, or the compressor in Radeland. In Figure 2.1 they are coloured blue. The first one is an element of the component LNG terminals, whereas the second one is an element of the component compressors. Hence, many elements make up a component. However, all elements are referring to different facilities by default. This means in a single network, one cannot have two elements of a component describing the same facility. The structure of elements is described below.

Attribute

"Attribute" is a term that is being used for the individual labels of the values that are associated with the elements. Examples for this term are gas "pipeline diameter", "maximum capacity", "max gas pipeline pressure", to name just a few and in Figure 2.1 they are coloured yellow. Overall there will be several hundred attributes in the SciGRID_gas project. However, the same attributes can occur in more than one component, e.g. "max flow capacity" exist for pipelines and also for compressors. Throughout the project, we have tried to keep the units of such attributes the same, so that there is no unit conversion required.

Attribute value

Each attribute has a value, most likely a number or a string. In Figure 2.1 they are coloured black. While boolean (*True*/*False*) is also allowed, more likely a "1" will stand for *True* and "0" for *False*. However, some attribute values might not be given in the data source, therefore a no value attribute value does exist. In the Python code it is termed *None*.

The Figure 2.1 depicts the relationships between the terms "gas data set", "component", "element", "attribute", and "attribute value". As can be seen, a single gas data set consists of several components, where each component contains several elements, and each element has several attributes, which each come with a value, where "None" stands for unknown value. The heuristic processes described in this document at a later stage will fill those "None" values with generated values.

Gas component types

A gas transmission network consists of different components, such as pipelines, compressors, etc. For the SciGRID_gas project a hand-full of components have been implemented, and will be described here briefly:

- *Nodes*: In a gas network, gas flows from one point to another point, which are given through their coordinates. All elements of all other components (such as compressor stations and power plants) have an associated node, which allows for the geo-referencing of each element. Overall the term "nodes" will be used throughout this document, as it aligns with graph theory aspects.
- *PipeLines*: *PipeLines* are one of the main components of the gas pipeline network. *PipeLines* allow for the transmission of the gas from one node to another. However, each pipe is unique. They might have different diameter, capacity or max pressure. In addition, a single *PipeLine* can connect several nodes. Therefore it could go from "Radeland" to "Bottrub" and then follow on to "Frankfurt". However, *PipeLines* do not need to connect more than 2 nodes, but can. The order of those nodes is important, and indicates the flow direction.
- *PipeSegments*: *PipeSegments* are almost identical to *PipeLines*, However, are only allowed to connect two nodes. Thus they have one start node and one end node, and are not passing via other nodes or other component elements in between, such as compressors or LNG terminals. Hence, any pipeline can easily be converted to multiple pipe-segments.
- *Compressors*: *Compressor* represent compressor stations, which are important. Gas travelling through the gas pipeline loses pressure due to friction on the pipeline walls and other factors. This will reduce the throughput of the gas amount. Hence, every so often (~ every 150 km), a compressor station is required, which increases the pressure of the gas, and Hence, allows the gas to flow through the gas pipeline. A gas compressor station contains several gas compressors units (turbines). Knowing the individual gas turbines is of an advantage, as those turbines can be combined in different ways, such as in series, or parallel, or combinations of those two options.
- *LNGs*: (LNG terminals and LNG storages) Some of the gas, which is being used throughout Europe, is supplied via ships to LNG terminals and LNG storage facilities. (From here onwards the acronym "*LNGs*" will be used instead "LNG terminals and LNG storage facilities".) As the transmission of gas would be extremely inefficient due to its volume, the gas state is changed to the liquid form (LNG gas), and then shipped. Ships arriving in Europe need special LNG terminals that can store LNG gas and subsequently re-gasify it. The storage and re-gasification of the gas are combined in the *LNGs* component and need to be part of any gas network for Europe.
- *Storages*: Part of the gas network will be gas storages. Gas storages are being used as gas pipeline capacities or gas production capacities might not be able to cover high demand periods, such as during the winter. Hence, large gas storage units are being filled during the summer periods while the overall demand is low, and if capacities of net supply allow it. This gas is then used during the winter period, and can compensate for shortcomings of the gas network or gas supply. Almost every country has their own gas storage units, ranging from smaller units to compensate for daily fluctuations to larger units, which compensate seasonal fluctuations. For the SciGRID_gas model the larger seasonal storage units are of more importance than the smaller ones,

as we are interested in the transmission gas pipeline network. However, any gas storage can be added and implemented into the gas network data set.

- *Consumers*: Part of the gas pipeline network is the knowledge of gas demand. Gas is added to the network at LNG terminals and European boundary cross border points. One type of users is the gas power plants. These can be added to the SciGRID_gas model, as this will specify local gas demand. In addition other consumers, such as city gas providers and large industries can also be added to the network data set.
- *Production*: These can be wells inside a country where gas is pumped out of the ground. Most of the gas used in Europe comes from outside of the EU, However, there are several smaller gas production sites scattered through Europe.
- *BorderPoints*: BorderPoints are cross border points (between different countries), which are mostly for the purpose of accounting the gas flow. Most large gas pipelines have cross border stations, e.g. Ellund (lat/long: 54.80181, 9.289079) at the border between Germany and Denmark, with gas facilities on both sides of the border.
- *EntryPoints*: These are special border points, as they are at the borders of the European Union and will be the gas entry points for the SciGRID_gas model data set.
- *InterConnectionPoints*: These are points between gas transmission operators, and will be found mainly within Europe, in particular at country borders. However, they can also be found within a single country, if there is more than one gas transmission operator.

Element structure

As described above, elements are describing individual facilities, such as compressors or LNG terminals. However, the overall structure of those elements is the same for all elements of all components. The overall structure of those elements is described in the following part:

- *id*: A string, that is the ID of the element, and must be unique.
- *name*: A string that is the name of the element, such as "Compressor Radeland".
- *source_id*: A list of strings that are the sources of the element. As several elements from different sources could have been combined in a single element, one might need to know which are the original ids of the original sources.
- *node_id*: This is the ID of a geo-referenced point to which an element of the network is associated to. For a compressor, this will be just a single node_id, However, for a gas pipeline, that starts at one point and finishes at a different point, this entry would be a list of at least two node_id values.
- *lat*: This is the latitude value of an element. For pipelines, lat is a list of latitude values if known. The geo-referenced projection of the element that is being used in the SciGRID_gas project is: World Geodetic system 1984 (epsg:4326).
- *long*: The longitude, analogue to lat.
- *country_code*: This is a string indicating the official 2-digit country code (Alpha-2 code, see Chapter 5.4 for list of countries and their code). It represents the location of the element. As pipelines can pass through more than one country, the country code for pipes is the list of country codes of the countries the pipeline is passing through.
- *comment*: This is an arbitrary comment that is associated with the element.
- *tags*: This dictionary is reserved for OpenStreetmap data. It contains all associated key:value-pairs of an OpenStreetMap item.

In addition, there are three further groups of attributes to each element. Throughout the SciGRID_gas project, they have been coded as "dictionaries". They are called:

- *param*
- *method*
- *uncertainty*.

The structure within each dictionary is the same, However, their meaning is different. First of all the dictionary *param* (short for "parameter") contains a list of attributes and their values. This list of attributes will be different for each component. For the component *PipeLines* they might be pipeline diameter, max pipeline pressure, and max pipeline capacity. For the component *Compressors* they might be , such as number of turbines, overall turbine power, energy source of turbine and more.

So the other two attribute dictionaries are *method* and *uncertainty*. Each of those two dictionaries contains exactly the same list of attributes as the "param" dictionary. However, their attribute values reflect the name of the dictionary. E.g. the attributes in the dictionary *method* contain the information on the method used to derive the attribute value that is stored in the param dictionary. Here methods of value generation can include heuristic methods names (in form of strings) that have been implemented in the SciGRID_gas project. However, if attribute values are not being generated by the SciGRID_gas project, but originate from one of the input data sources, then the attribute values in the *method* dictionary is set to "raw".

Similar is the content of the *uncertainty* dictionary. It contains information on the uncertainty of the attributes from the *param* dictionary of that component. Again all attributes listed in the *param* dictionary are also present in the *uncertainty* dictionary. The attribute values here reflect the uncertainty of the attribute. Here, it is assumed, that attributes with a method of "raw" have an uncertainty of zero. Only for those attributes, which were generated during heuristic SciGRID_gas methods an uncertainty larger than zero will be specified.

2.2 Summary

The SciGRID_gas software is designed to construct a gas transmission network data set form different open source gas component data sets. The gas transmission data set needs to be available and stored in a precise and predefined way, which was described in this section. We have identified several *component*-types of a gas transmission network grid, like pipelines, compressor stations, LNG-terminals, etc. Each specific facility that falls under such a component is considered an *element* of that component. Each element is described by a list of *attributes* and correspondent *attribute values*.

CHAPTER

THREE

DATA SOURCES

Two thirds of the gas used in Europe is imported from non-EU states, and all gas required for the consumption needs to be distributed through the existing gas transmission pipelines in Europe. In the future gas consumption might rise, leading to additional pressure on the current infrastructure. In addition, gas facilities could play a vital role in reducing CO_2 emission, as excess electricity could be converted to gas, that could be stored and transmitted throughout Europe with the existing gas network. Hence, a reliable network data set for the European transmission network is essential. The data required for such models ranges from pipeline diameter, gas pressure within the pipeline, actual pipeline length, pipeline capacity, and underground storage volume to name just a few.

However, such data is the property of the transmission system operators (TSOs) and is therefore generally not freely available in the form and depth that is required for modelling purposes. The major reason for the difficulty of obtaining of such data is that most of the gas network infrastructure, namely pipelines, is buried underground. Thus a pipeline diameter is hard to estimate locally. In addition, almost all of the data is commercially sensitive.

However, there is a public drive to gather such data and subsequently make it available. The major platform through which this is occurring is the Open Street Map database [Hel18]. OSM is a geo-referenced database through which people can supply geo-referenced information on all man-made and natural structures, ranging from mountains to buildings. To achieve this, people throughout the world wander the globe and geo-reference everything that they can find. This also includes gas-pipeline markers, compressor stations or LNG terminals. However, the major problem remains that one cannot measure or estimate the diameter of the underground pipelines, or the number and size of the compressor turbines, as compressors are within buildings, which are fenced off. Hence, such information is hardly supplied to the OSM platform.

Nevertheless, some data is made available by gas transmission network operators, through different channels. E.g. information on the size and number of compressors could be made public through a press release, as part of a refurbishment. An example is given below (https://www.maz-online.de/Lokales/Teltow-Flaeming/Neue-Verdichterstation-entsteht-in-Radeland):

> "Die Eugal-Pipeline dient dazu, Gas aus der neuen Ostseepipeline Nord Stream 2 bis zur tschechischen Grenze zu leiten. 275 Kilometer von ihr verlaufen in Brandenburg. Grundsätzlich soll die neue Leitung parallel zur bestehenden Opal-Pipeline gebaut werden."

In addition some information can be found on company web pages, (https://www.open-grid-europe.com/cps/rde/SID-752BB6B5-E0A975F2/oge-internet-preview/hs.xsl/NewsDetail.htm?rdeLocaleAttr=en&newsId=50190C3B-E14F-4685-9E64-E40EEAB57A28):

> Open Grid Europe (OGE) is investing roughly EUR 150 million at its compressor station in Werne to improve the security and flexibility of energy supply for North Rhine-Westphalia and Germany. The upgrade of the station, which is one of the hubs of the pipeline network, will allow gas flows to be switched (reversed) from north to south and south to north. In addition, OGE is preparing the station for the upcoming transition from L- to H-gas. Through this fitness programme, the station's transmission capacity will increase by about 500,000 to 6.5 million m^3/h, which is equivalent to the annual consumption of more than 2,100 single-family homes. The project, which is due for completion at the end of 2018, is fully on track."

The data available can be separated into two different groups:

- OSM data: Data can be found in the OSM data base. OSM data is well geo-referenced, but contains little meta-information (information on the facility attributes, such as pipeline diameter or pipeline capacity). OSM data is very helpful to obtain accurate routes of pipelines.
- Non-OSM data: Non-OSM data have in general lower geographical accuracy but contain a lot of meta-information. Unfortunately, such information is only known for a few facilities. One exception to this rule are shapefiles from TSOs. They are rare, but well geo-referenced. However, the resolution of the meta information can vary from TSO to TSO.

One of the main challenges for SciGRID_gas is that, gas transmission data is incomplete and accumulated from different sources. Also such different sources can have different properties for one and the same facility. Hence, it is important to know, which data set supplies which information. Hence, this chapter here will introduce the relevant data sets (e.g. INET), starting off with the components, the elements for each component and then the attributes for each element.

3.1 Non-OSM data

Non-OSM data includes data from internet research, TSO press releases, TSO transparency platform, TSO public data, national open-source gas network data sets[1], etc.

Some of the TSO information had to be made available due to EU-regulations. Other information has been made public as part of a company's self presentation and advertisement. The information used by the SciGRID_gas project focuses on:

- the quality of the data
- the format of the data
- the level of representation of the data
- and the copyright restrictions on the data.

In addition, each data source is unique. Source specific tools need to be developed, so that all data sources can be made accessible for the SciGRID_gas project in the format as described in later chapter releases.

A significant portion of the project was spent on finding non-OSM data sets . Further data sources might be available, but unknown to the authors. If the authors are made aware of additional sources, the project will try to incorporate those, as this would only increase the depth of the data available and increase the applicability of the gas network data set and model.

Non-OSM data sources are very specific, addressing only certain aspects of the entire gas infrastructure. E.g. the GIE [GasIEurop20] data set supplies information on the daily gas flow in and out of gas storages in LNG terminals. However, they fall short on specifying the fundamental information of the actual physical location. Other data sets, such as the LKD [FMWP+17] data set is quite detailed in respect of pipelines, compressors and consumptions, however, only available for Germany.

Hence, the main task is to look closely at each data source, distil which data attribute values can be used, how it can be downloaded and incorporated into our SciGRID_gas model, and identify the copyright restrictions on the data source.

Due to copyright regulations, there are roughly two groups of data:

- Non copyright restrictive data (N-CRRD): here the copyright does not restrict the download, use and distribution of the data.
- Copyright restrictive data (CRRD): here the data can be downloaded and used internally, but not re-distributed to others.

[1] An entire gas network data set is only available from the UK, see https://www.nationalgridgas.com/land-and-assets/network-route-maps'.

The following is a list of the data sources that will be used throughout the project and an indication into which group of copyright restriction they fall:

- **OSM** (https://www.openstreetmap.org) (N-CRRD)
- **GB** (https://www.nationalgridgas.com/land-and-assets/network-route-maps) (CRRD)
- **NO** (https://www.npd.no/en/about-us/information-services/available-data/map-services/) (N-CRRD)
- **LKD** (https://tu-dresden.de/bu/wirtschaft/ee2/forschung/projekte/lkd-eu) (N-CRRD)
- **ENTSOG** (https://transparency.entsog.eu/) (CRRD)
- **EMap** (https://www.entsog.eu/sites/default/files/2020-01/ENTSOG_CAP_2019_A0_1189x841_FULL_401.pdf) (CRRD)
- **GIE** (https://www.gie.eu/) (N-CRRD)
- **GSE** (https://www.gie.eu/index.php/gie-publications/databases/storage-database) (N-CRRD)
- **IGU** (https://www.igu.org/) (CRRD)
- **GasLib** (http://gaslib.zib.de/) (N-CRRD)
- **INET** (see `Refs_InternetData`) (N-CRRD).

Each data set and source comes with a different copyright regulation. The copyright can be rather non-restrictive (e.g. INET) or can be restrictive (IGU). It is attempted to use only freely available data, so that such data can be re-distributed. In more restrictive data cases (IGU, GB), it is not allowed to download the data and distribute it to others. However, it is allowed to let other potential users know of the location of such data and supply them with tools, that allow them to carry out the same data download and subsequent incorporation of the data into a gas network data set.

Note:

In case that other users are aware of other data sources, that might be useful to this project, please get in touch and supply us with a brief description of the data and the location of such data, so that additional tools can be developed to incorporate the data in this project. Please use the following email address: developers.gas(at)scigrid.de

3.2 EntsoG-Map (EMap) data set

This section contains information on the content and nature of the so called **EntsoG-Map** (**EMap**) data set, how this data was generated, its format, and its content.

3.2.1 Overview of the EMap data set

The gas pipeline and gas facility from the EMap data set is of great importance to the SciGRID_gas project. It is one of a few data sets available that supply geo-referenced pipelines, storages and production facilities. In addition, it contains some attribute values in respect of gas pipelines that are fundamental for the gas data model. This data set was generated by converting the freely accessible PDF map of the European gas network into the SciGRID_gas data structure, where original approach was proposed by Yueksel-Erguen and Zittel [YEZ20]. Table 3.1 summarises the number of elements for each component found:

Table 3.1: EMap component summary

Component Name	Count
BorderPionts	0
Compressors	0
ConnectionPoints	0
Consumers	0
EntryPoints	0
InterConnectionPoints	0
LNGs	0
Nodes	2069
PipeSegments	3088
Production	117
Storages	238

Overall, this current data set consists of 216239 km of pipelines. In addition, a map (see Figure 3.1) visualizes the data for Europe.

Warning: Due to the nature of the generation of the EMap data set, as will become clearer when reading the data generation sub-section, the actual accuracy of the elements location is limited, and it is assumed that the "true" locations of nodes and pipelines could be out by ~20 km and up to 100 km.

3.2.2 Origin of the data

The origin of the EMap data is a map in PDF format supplied by EntsoG. EntsoG is the acronym for "European Network of Transmission System Operators for Gas", and is an association of the European transmission system operators.

The EntsoG map covers all of Europe, including the non-EU states Russia, Ukraine, Belarus, Georgia, Aserbidian, and others for the energy source gas. This map is being published on an irregular basis, and the latest version is from 2019. The project SciGRID_gas is very fortunate, that a map version of the gas pipelines, drilling platforms and storage facilities is available. As part of the project, tools have been created to incorporate some of the information from the map into the project.

The latest map version of EntsoG is available from the following link: https://www.entsog.eu/sites/default/files/2020-01/ENTSOG_CAP_2019_A0_1189x841_FULL_401.pdf

The EntsoG map is freely available as a PDF file. Several steps need to be carried out to convert the PDF into the SciGRID_gas data structure. For this several Python tools have been created. However, this process cannot be fully

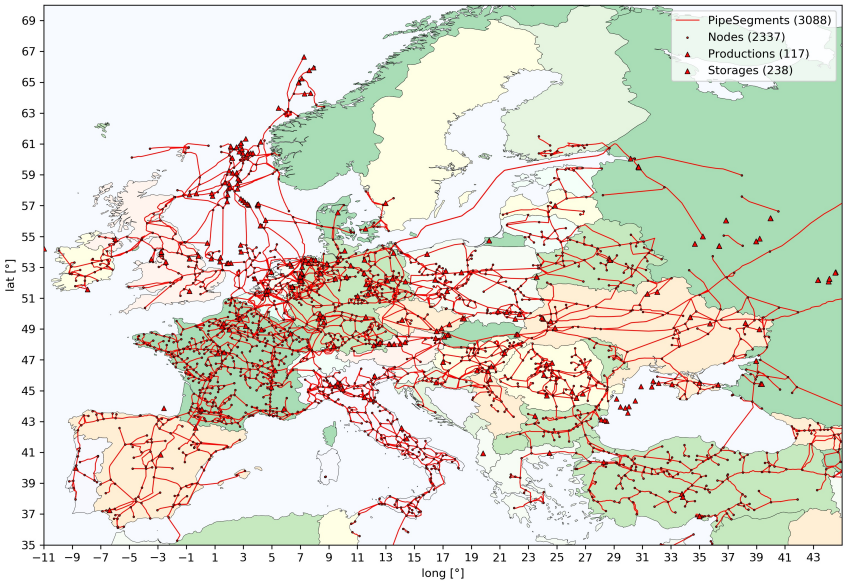

Figure 3.1: Map of components of the EMap data set.

automated, but all steps have been taken to automate as many aspects as possible, whereas some cleaning up will need to be carried out by hand by the user. The process of generating the data set is being described in more detail in Chapter 3.2.5, whereas the following section will inform the user of the data density of all important attributes.

3.2.3 EMap data density

Here the "data density" of EMap data will be presented. This is important so that one can get a better understanding of the data. Here "data density" is defined as follow: this is the ratio of the number of usable (not missing) attribute values over the number of elements of the component. This is explained through an example: supposedly the EMap would have two *Storages* elements. And one of the facilities has supplied a storage volume, whereas the other one does not. Hence, the data density would be 50 % for the attribute storage volume. For each component the data density for the most relevant attributes will be given next.

PipeSegments

Overall there are 5146 *PipeSegments* elements in the resulting EMap data set. Table 3.2 summarizes the data densities for the most important PipeSegments attributes:

Table 3.2: PipeSegments data density

Attribute name	Data density [%]
length_km	100
pipe_class_EMap	94.6
exact	100

As each element of the component *PipeSegments* originated from a line in the original PDF map, and as this map was geo-referenced, a length for each element could easily be determined, Hence, the overall data density for the attribute

"length_km" is 100. In addition, for the attribute "exact" a blanket value of 3 has been assumed, indicating that the topological accuracy could be out between 10 and 100 km. The "pipe_class_EMap" is a value that was generated as part of the data generation process. The original PDF file contained three different layers for three different pipeline thicknesses: "small", "medium", and "large". This was given for all pipelines of the map, except for Germany. For Germany all pipelines, independent of their thicknesses, were grouped into a single layer. Hence, the conversion process was not able to differentiate between "small", "medium", or "large", and Hence, no value was given to any pipelines in Germany. Therefore the overall density of the attribute "pipe_class_EMap" is 94.6 %.

Storages

Overall there are 238 *Storages* elements in the resulting EMap data set.

The extraction process was not able to retrieve any further information for the *Storages*, except their locations.

Productions

Overall there are 117 *Productions* sites in the resulting EMap data set.

Again, The extraction process was not able to retrieve any further information for the *Productions*, except their locations.

Nodes

Overall there are 4323 *Nodes* elements in the resulting EMap data set.

Table 3.3 summarizes the data densities for the most important *Nodes* attributes:

Table 3.3: Nodes data density

Attribute name	Data density [%]
exact	100
elevation_m	100

Here again, the information that has been used to generate a value for "exact" is the same as applied for the *PipeSegments*. Hence, each Nodes element has a value of three.

The elevation attribute "elevation_m" was not retrieved from the EntsoG PDF map, but was generated using an API from Bing.

3.2.4 Copyright

Copyright

Based on the legal framework, all of the EMap was generated in such a way, that it has a copyright that does not restrict us from making the data available to other users.

Hence, the following applies to the EMap data:

Disclaimer

The EMap data set is supplied on a best-effort basis only, using available information as documented gathered from the WWW. While every effort is made to make sure the information is accurate and up-to-date, we do not accept any liability for any direct, indirect, or consequential loss or damage of any nature–However, caused–which may be sustained as a result of reliance upon such information.

3.2.5 EMap generation processes

In this part of the chapter, a description is being given, on how the data set was generated, originating from a PDF document and ending up as a SciGRID_gas data object. Lustenberger et al. [LSS+19] presented a similar pathway of generating the same data set, however using the non-open ArgGIS tool. Here, the open source tool QGis is being used. A comparison of those to data sets will be carried out at a later stage.

Below is a general overview of the steps that have been implemented to convert the EntsoG PDF map into a single SciGRID_gas gas network data set:

- Separate the individual layers from the original PDF map into separate files (**PDF Layer generation**).
- Convert the above PDF files into highest resolution TIFF files (**PDF to TIFF conversion**).
- Geo-reference the TIFF file, which will produce raster layers (**Geo-reference of TIFF files**).
- Convert the raster layers into SciGRID_gas *PipeLines*, *Storages* and *Productions*, for each country individually (**Generation of country specific SciGRID_gas network elements**).
- Remove little pipelines that are wrong artefacts of the PDF to TIFF conversion process, for each country individually (**Removing wrong elements**).
- Joining the above components by country, Hence, resulting in component data sets (**Combining country specific data sets**).
- Joining above data set into a single SciGRID_gas network data set, which will consist of many un-connected *PipeLines*, *Storages* and *Productions* (**Joining data**).
- Joining lose *PipeLines*, *Storages* and *Productions* to form one single SciGRID_gas network (**Generation of a single coherent SciGRID_gas data set**).

The overall outcome of this process is the conversion of the PDF map into more than 3000 *PipeLines* elements, more than 200 *Storages* elements and more than 100 *Productions* elements throughout Europe, including Russia, and other non-EU states, resulting in a total length of more than 200000 km of pipe-lines.

The steps of how to convert the PDF map into the SciGRID_gas data are presented below.

PDF Layer generation

The data source is a PDF map of the European gas transmission network, including sites of *Productions*, *Storages*, under-sea *PipeLines* and overland *PipeLines* in different thicknesses, based on their throughput. This map can be downloaded from the EntsoG web page (see link above). As the PDF consists of several layers, one can use an external tool to separate those layers and remove unwanted layers, such as legend, coastal lines, or gas fields. This process needs to be done by hand in an application, such as "Adobe Acrobat Reader". In this application, the layers tool can be selected, and individual layers can be saved as individual PDF layers. Below (see Figure 3.2) a screen shot shows the "Adobe Acrobat Reader" software with the EntsoG map loaded, and the layers tab expanded. Several layers can be seen in the screenshot and are part of the EntsoG map, such as "CAPDATA", "datapanel GRAY", ">>>LEGEND" etc.

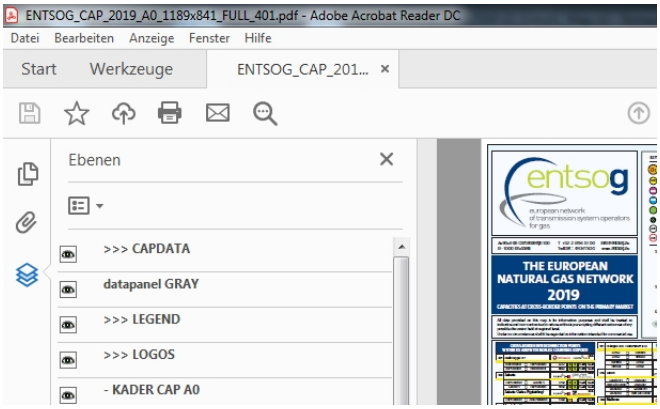

Figure 3.2: Screenshot of "Adobe Acrobat Reader" with the expanded layers tab, and a list of some layers to the left.

Most of the layers that are present in the EntsoG map do not contain information that is needed for this project, and can be discarded. But the following layers are required for the SciGRID_gas project:

- ">> STORAGE NONEU", will be part of the SciGRID_gas "Storages" component
- ">> STORAGE TYNDP", will be part of the SciGRID_gas "Storages" component
- "= DRILLPLATFORMS =", will be part of the SciGRID_gas "Productions" component
- "PIPELINES_NEW_GERMANY", will be part of the SciGRID_gas "PipeLines" component
- "PEPELINES > SMALL", will be part of the SciGRID_gas "PipeLines" component
- "PEPELINES > MEDIUM", will be part of the SciGRID_gas "PipeLines" component
- "PEPELINES > LARGE", will be part of the SciGRID_gas "PipeLines" component
- "=== NORTHSEA - pipes > GAS", will be part of the SciGRID_gas "PipeLines" component
- "=== NORTHSEA - pipes > CONDENSATE", will be part of the SciGRID_gas "PipeLines" component.

These layers need to be exported individually into single PDF files.

In addition, further layers are needed for the geo-referencing process at a later stage: - "BORDERS" - "SHORES" - "LANDMASS"

All those three layers need to be exported combined into a single PDF file, which will be referred to as the layer "*ENTSOG_Borders*".

The data is stored in the following folder:

"../SciGRID_gas/Eingabe/Maps/EntsoG_2019/01_PDF/"

PDF to TIFF conversion

For later processes, the PDF files need to be converted into TIFF format. For this an external application, such as https://onlineconvertfree.com/de/convert-format/pdf-to-tiff/, can be used. The user should select an application, which retains as much resolution as possible.

Resulting TIFF files were stored in the following folder: "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/02_TIFF/"

Geo-reference of TIFF files

As the projection of the original map is unknown, we need to determine the projection using an external application, such as **QGIS**. For this one needs to load the layer *ENTSOG_Borders* from above. The overall plan is to geo-reference this layer *ENTSOG_Borders* and in a second step apply the determined geo-referencing to all the other selected layers.

Hence, one needs to load the layer *ENTSOG_Borders* into **QGIS**, which is not geo-referenced at this stage. In addition, one needs to load a geo-referenced layer of Europe or of the area of interest as well. Care needs to be taken, that the second layer is projected in the projection that has been selected for the SciGRID_gas project. In the case for Europe, the projection "epsg:4326" was selected.

Now the following **QGIS** process is required: "Georeference GDAL". This is a plugin, that can be installed from within **QGIS**, for **QGIS** versions of smaller than 3. For version 3.0 and newer, this plugin comes pre-installed with the base installation. (If you have problems finding "Georeference GDAL" in **QGIS** 3.x, then follow instructions under link https://gis.stackexchange.com/questions/274503/georeferencing-in-qgis-3-0).

The tool "Georeference GDAL" can be found under "Raster" and then "Georeferencer. . . ".

Here for SciGRID_gas the following steps need to be taken:

- Open **QGIS**.
- Open a layer of the European country layers, here the user can use the "TM_WORLD_BORDERS-0.3" layer [San19], that can be downloaded from the following site: https://koordinates.com/layer/7354-tm-world-borders-03/.
- Start the Georeferencer, and new Georeferencer window will open.
- Press the [Open Raster] icon, and select the layer *ENTSOG_Borders*.
- Open the "Transformation Settings" window by pressing the [Transformation Settings] icon, and select the following as depicted in Figure 3.3:

Here, the user needs to select the following:

- "Transformation type": "Thin Plate Spline"
- "Resemble method": "Cubic Spline"
- "Target SRS": "EPSG:4326 - WGS 84"
- "Output raster": "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/03_Raster/ENTSOG_Borders.tiff"

and press the [OK] button to finish off this setup.

- In the "Coordinate Reference System Selector" select the "epsg:4326" coordinate reference system.
- Select the [Add Point] icon.
- By pressing the [Shift] button and use the mouse wheel, the user can find striking features on the TIFF map and select the location by pressing the left mouse button on top of it. Here as an example (see Figure 3.4) the border between Russia and Poland is being displayed, and a good location would be where the boarder meets the Baltic Sea.
- After pressing the left mouse button, the following window will appear (see Figure 3.5):

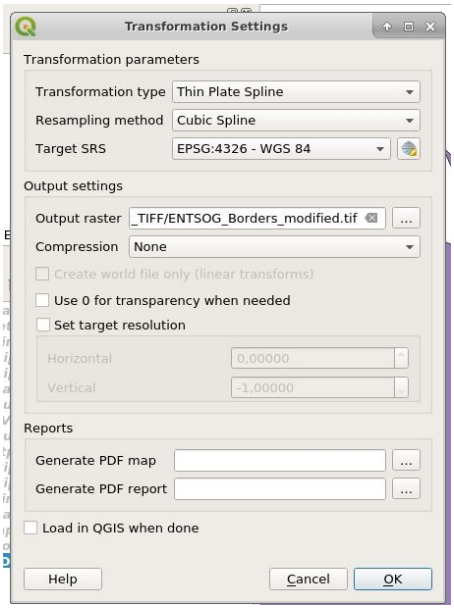

Figure 3.3: Screenshot of the "Transformation Settings" window.

Figure 3.4: Screenshot of a sample location of the Russian-Polish border in the Gdansk Bay.

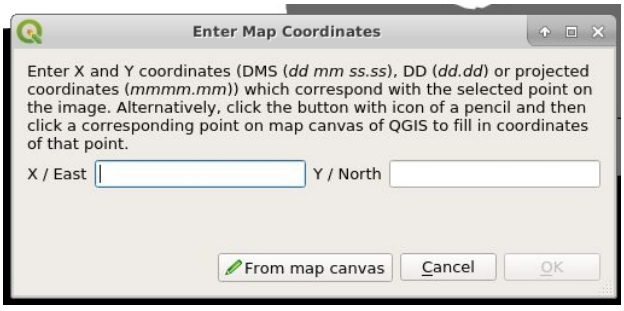

Figure 3.5: Screenshot of the new window "Enter Map Coordinates".

Here the user needs to press the [From map canvas] button.

- Find on the loaded georeferenced map (TM_WORLD_BORDERS) the appropriate location and press the left mouse button again. This will populate the X/Y coordinates in the "Enter Map Coordinates" window, as shown in Figure 3.6.

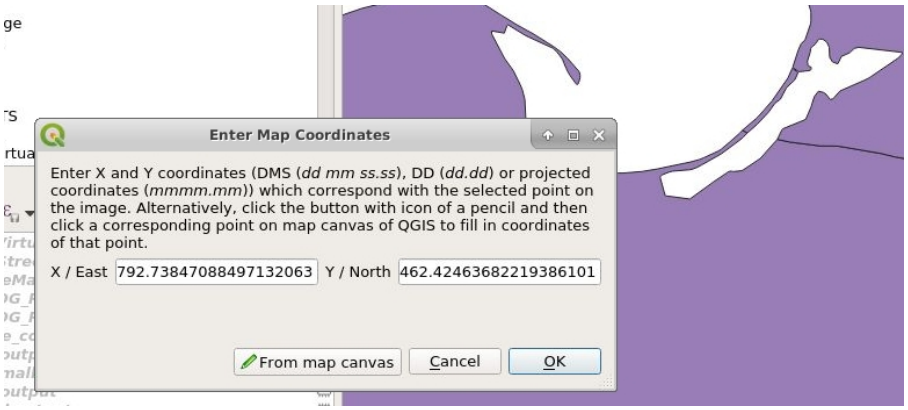

Figure 3.6: Screenshot of the window "Enter Map Coordinates" with the populated X/Y values.

Press the return button to lock in this geo-referenced pair of values. - In the "Georeferencer" window an entry should appear in the "GCP table", where the table is located below the map (Figure 3.7).

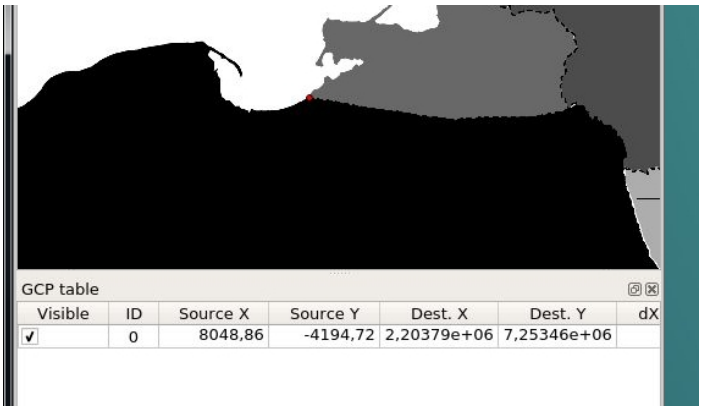

Figure 3.7: Screenshot of the "GCP table" entry with the new pair of coordinates, within the "Enter Map Coordinates" window.

- Repeat this process for a large number of points throughout Europe. Select points on the peripheries of Europe, but also select points within Europe, e.g. the three border location of Belgium, Germany and the Nederland, or other territorial and geographical features, such as Isle of Guernsey or Isles of Scilly AONB. However, try not to use too many point pairs, a good spread is more important. (Here about 200 points were selected in the original process.)
- Now the user can check the geo-referencing by pressing the [Start Georeferencing] icon. This process might take several minutes. It will result in a new layer in the QGIS Layers list. Try to visualize this new layer and the "TM_WORLD_BORDERS-0.3", by setting one layer slightly transparent, so that one can eye up the newly projected layer "ENTSOG_Borders" with the "TM_WORLD_BORDERS-0.3" layer and look for areas of large difference (example given in Figure 3.8). Now more pair points can be added to rectify areas of imperfect geo-referencing, until the user is satisfied with the result. As an example, Luxembourg is presented here, and one can see that the locations of the 3 Borderpoints of Luxembourg, the Netherlands and Germany on the north and Luxembourg, the Netherlands and Belgium on the west are not perfect. Hence, placing additional geo-referencing pairs might help to rectify this discrepancy.

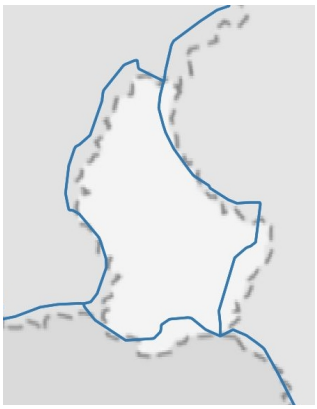

Figure 3.8: Screenshot of both layers around Luxembourg, showing the mismatch of the projection.

- If the user is satisfied with the geo-referencing and the underlying pairs of values, the user needs to save the point pairs, as they will be used for the other TIFF layers. This can be achieved by pressing the [Save GCP Points as] icon in the "Georeferencer - ..." window. A window will pop up and the user will need to enter a location and a file name of the points table.
- **Now the user needs to apply this same geo-referencing to the other layers of the EntsoG map. For this carry out the follow**
  - Select a new TIFF file from the layer list above in the "Georeferencer - ...".
  - Open the previously saved GCP table by pressing the [Load GCP Points] icon.
  - Select a different destination file under the [Transformation Settings] window.
  - Initiate the geo-referencing process by pressing the [Start Georeferencing] icon.

This should be carried out for the *PipeLines*, *Storages* and *Productions* layers.

The overall output will be that the user will have created raster TIFF layers of the EntsoG gas elements, such as *PipeLines*, *Storages* and *Productions*, which are geo-referenced. Those resulting files should be stored in the folder: "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/03_Raster/".

Generation of country specific SciGRID_gas network elements

In the next step, the user needs to convert the raster layer into SciGRID_gas elements. For this Python code can be executed by the user.

The Python routines are combined in the **M_Maps** module, and can be access with the **M_Maps.read()** function. A large list of settings are required, However, they have been implemented into the code as default values, if no other values are supplied.

The previous steps created a geo-reference raster layer. However, this raster layer needs to be converted into polygons, which subsequently need to be converted into SciGRID_gas *PipeLines*, *Storages* and *Productions*.

The functions that have been developed to carry out those transformations are listed below for each sub-section.

Raster to polygons

The main function that is being used to convert the raster files into polygons is called **M_Maps.raster2Polygon**(). This function uses the freely available **GDAL** Python module that can be downloaded and installed for Python. The resulting file format is of type shapefile, and resulting files will be stored in the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/04_Polygon/".

The above process created a very large number of polygons, where some of the polygons are of the size of the PDF raster scanning resolution. To reduce the number of polygons, horizontally adjacent polygons are combined into single polygons, reducing the number of polygons by about 25 %. Results of this process are written as shapefiles into the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/05_Polygon/".

Manual shapefile clean up processes

After the above step the user needs to carry out a manual process. This is required, as the polygons created contain spatial "mistakes", as polylines are surrounded by a polygons, as can be seen in Figure 3.9.

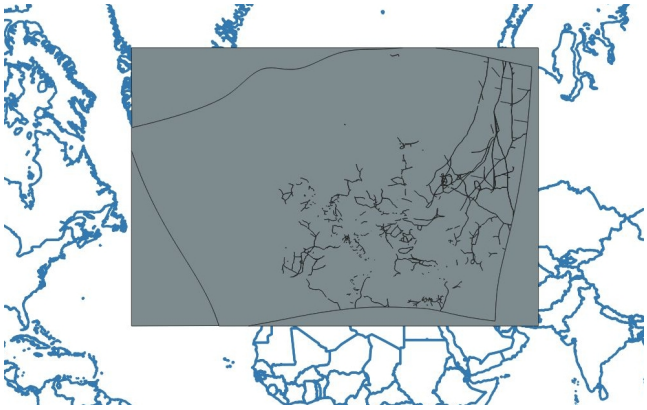

Figure 3.9: Sample shapefile, prior to clean up, where entire shapefile area is covered by one or several large polygons.

The goal is to remove all those polygons that are not lines, as is the case in Figure 3.10.

Figure 3.10: Sample shapefile, after the clean-up, where all polygons are pipelines.

This can be achieved by using an application, such as **QGIS**, and selecting and removing the unwanted polygons. Figure 3.11 shows the entire shapefile, where a single polygon has been selected (yellow) which has been removed in the next process step, resulting in Figure 3.12.

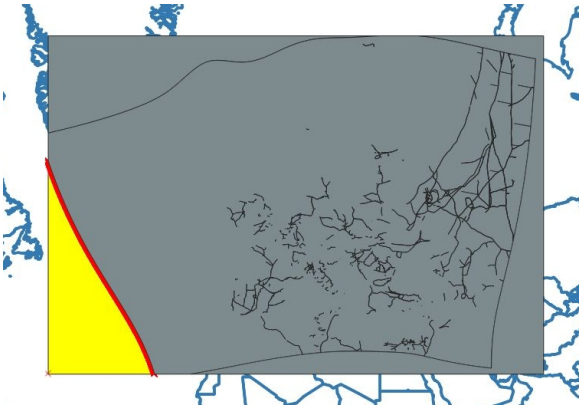

Figure 3.11: Sample shapefile, where a single polygon has been selected (yellow area with red stars)

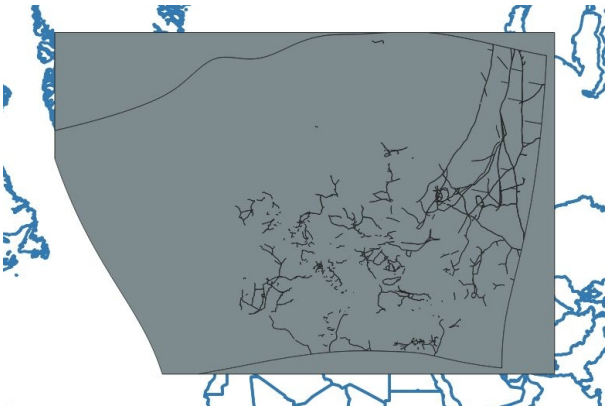

Figure 3.12: Sample shapefile, after the removal of the above selected polygon.

As can be seen in Figure 3.13, even areas between pipelines can be polygons (grey area between pipelines). These need to be removed as well, and have been selected as shown in Figure 3.14, and results of the removal process can be found in Figure 3.15.

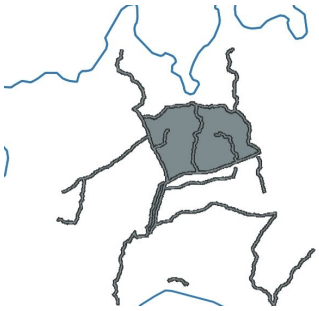

Figure 3.13: Sample shapefile, with polygon between pipelines.

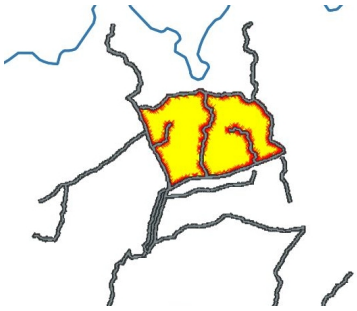

Figure 3.14: Sample shapefile, with polygon selected between pipelines.

Figure 3.15: Sample shapefile, with above selected polygon removed.

However, there are polygons between parallel lines of pipelines that need to be removed as well. Such a polygon between two parallel lines can be seen in Figure 3.16, which has been selected already. After the removal process (Figure 3.17) the two parallel lines are better visible and will make it easier for subsequent processes to carry out the conversion process from polygons to SciGRID_gas elements.

However, the resulting number of polygons per component group was still to large for any subsequent processes. Hence, the shapefiles were cut into country specific shapefile. Resulting files are written into the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/06_Polygon_CC/". Depending on the country/continent you want to create SciGRID_gas for, this could contain more than 1000 files. However, those files are small and partially contain no data e.g. there are no *Productions* sites in Luxembourg, but a set of shapefiles were generated for this component. However, the smaller file sizes allow for a faster processing of the EntsoG map data overall, as distance calculation, which form a large part of the ongoing processes, are being carried out with much smaller data sets.

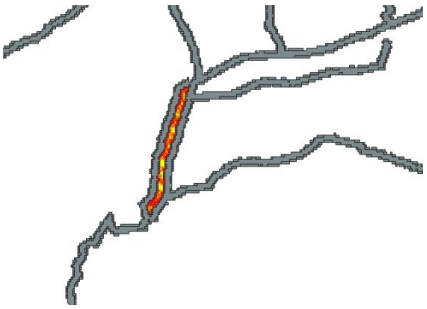

Figure 3.16: Sample shapefile, with polygon between two parallel pipelines selected (yellow and red).

Figure 3.17: Sample shapefile, with polygon between two parallel pipelines removed.

Polygons to SciGRID_gas elements

The main function that is being used to convert the polygons files into polygons is called **M_Maps.polygons2Netz()**. This function calls several functions from other modules, e.g. **GDP.GeoSerie**, or creates instances from other class definitions, e.g. **geometry.Centerline()**. For this processes to work, component specific parameters had to be determined, and will be part of the default settings. The outputs of this process are country and component specific SciGRID_gas component data sets.

Resulting data is being written into the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/07_A_CSV_CC/".

Here the resulting pipeline data sets received an attribute called "pipe_class_EMap". Entry values can be as follow:

- '1': The pipelines originated from the layer "ENTSOG_PipeLines_Large"
- '2': The pipelines originated from the layer "ENTSOG_PipeLines_Medium"
- '3': The pipelines originated from the layer "ENTSOG_PipeLines_Small".

Besides pipeline length, which will be generated dynamically at a later stage, this is the only attribute that was able to be extracted from the PDF map.

Removing wrong elements

During the digitization process and the subsequent processes of converting the data into a SciGRID_gas data set, wrong lines started to appear in the data sets. These need to be removed, as otherwise, they would be leading to wrong *PipeLines* elements, *Productions* sites, or *Storages* facilities. Hence, a function was written that removes pipe-lines, that are connected at only one end, and is called **M_Maps.multi_removeStichPipeLines()**. It was found that for some component elements, e.g. of type *PipeLines*, this function needed to be executed several times with varying settings, whereas it was not applied to any element of type *Productions*.

Resulting data is being written into the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/07_B_CSV_CC/".

A further function was designed that removes *PipeLines* elements, that are not connected at all. These are so called lone pipe-lines and can be removed by the function **M_Maps.removeLonePipeLines()**. Here a threshold of 2.55 km has been selected.

Resulting data is being written into the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/07_C_CSVs_CC/".

Combining country specific data sets

At this stage the overall data quantities have been reduced to such a level, that data combining processes can be applied. Hence, at a first step, the data file separation based on countries has been removed. The function **M_Maps.joinNetze()** joins all country data sets into a single component data sets. Resulting data is being written into the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/08_CSV/".

Joining data

Hence, to further reduce the number of *PipeLines* elements, the function **M_Maps.joinPipeSegments()** has been created. It joins pipelines together, in case that the following is given:

- Only two *PipeLines* elements are connected to this one node.
- No other other component elements, such as *Storages*, is connected.
- The attributes of those *PipeLines* elements need to be the same, e.g. large pipelines.

Resulting data is being written into the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/09_RawData/".

Generation of a single SciGRID_gas data set

In this section function are being described, that combine the different component elements into a single SciGRID_gas data set, and apply further functions on this data set to create additional connections, e.g. landbased pipelines and offshore drilling platforms.

Hence, the first function is **M_Maps.joinDataSets()**, which joins all the separate data sets into a single SciGRID_gas data set. Resulting data is being written into the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/10_Final/".

In a next step, the nodes of the SciGRID_gas data set are re-generated from all *PipeLines* elements, all *Storages* facilities and all *Productions* sites. Resulting data is being written into the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/11_Simpel/".

However, the resulting data set still consists of a large number of unconnected elements of type *PipeLines*, *Productions*, and *Storages*, that are not connected with the network. Hence, an additional function **M_Maps.spawningTree()** was implemented, using the spawning tree [GH85] method of connecting un-connected elements. Resulting data is being written into the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/13_Spawned_CC/".

For this process to work, the data needed to be "converted" with a process name **M_Maps.Pipes2Chunks_CC**(). This process interpolates the pipelines into chunks of 5 km length, where resulting data is being written into the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/12_Chunks_CC/".

The final function **M_Maps.joinElements()** creates a single SciGRID_gas data set, where all elements are connected into a single SciGRID_gas data set. Resulting data is being written into the folder "../SciGRID_gas/Eingabe/Maps/EntsoG_2019/15_Final/".

3.2.6 Summary EMAP Data

The data set was available through the internet and was downloadable as a PDF map from the "EntsoG" umbrella organisation. Tools have been created to convert the PDF into the SciGRID_gas data structure and make the data accessible throughout the SciGRID_gas project.

The Table 3.4 summarises the number of elements for each component found:

Table 3.4: Component element summary

Component Name	Count
BorderPionts	0
Compressors	0
ConnectionPoints	0
Consumers	0
EntryPoints	0
InterConnectionPoints	0
LNGs	0
Nodes	4323
PipeSegments	5146
Production	117
Storages	238

Below the current version of the EMap data set, after having applied all above steps, is presented for all of Europe in Figure 3.18, resulting in a total of 216239 km of *PipeSegments* elements.

3.3 Data summary

SciGRID_gas is based on open source data. To generate a gas pipeline network data set, one needs to access different data sets that were found throughout the project and presented here. Emphasis was given to depict the number of elements per component and the data density for each data set.

3.4 Summary

Gas component data sets come in different forms, licenses, formats and detail. The SciGRID_gas project can process such data and combine them to a consistent and reliable network data set.

The underlying gas component data sets were categorized into two different groups:

- OSM data: This is data originating from the OSM data base, containing well geo-referenced locations of gas facilities, such as pipe locations or gas storage facilities. However, it comes with very few meta information.
- Non-OSM data: These are all other data sources, which can "supply" detailed information on some of the gas facilities attributes. However, this information is sparse, as published only for a few facilities. Here, the INET

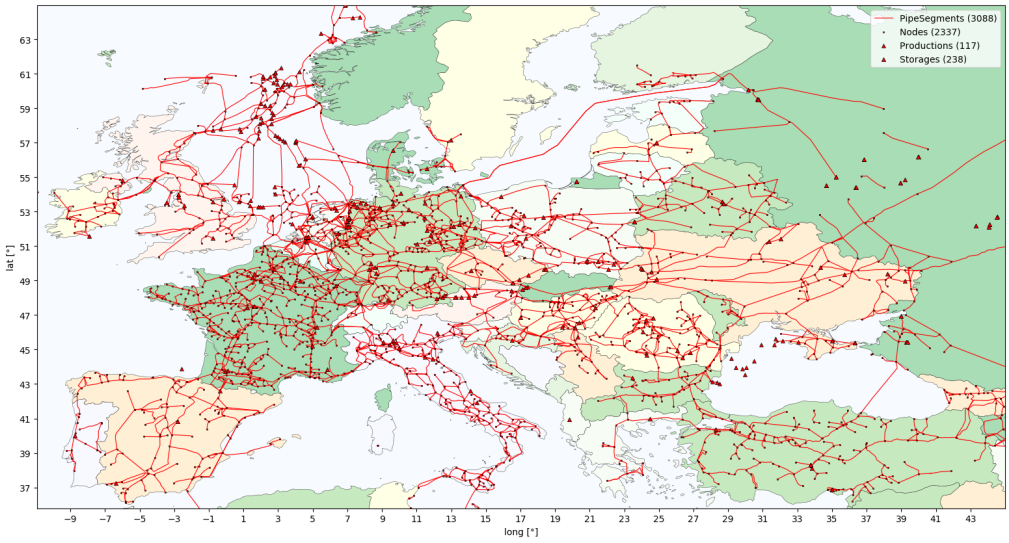

Figure 3.18: The pipelines, storage facilities and production sites of the EMap data set.

data set was introduced as an example of the non-OSM data set, and the pathway of converting the raw data from the www into SciGRID_gas project component structure.

Here detailed information on one or several data sources have been given, and should be used as a reference for later data processes.

CHAPTER

FOUR

CONCLUSION

This document here is the documentation of one of the data sets that is part of the SciGRID_gas project. This document here started off with the introduction of the SciGRID_gas project, such as funding, duration and goals. In a subsequent chapter the data structure within the SciGRID_gas project was described, such as components, elements, attributes and attribute values, so that the transmission data set could be an input to certain gas flow model. The third chapter introduced the EMAP data set, which is a data set that was generated by disecting a PDF map of the european gas network that is made available from EntsoG. The data presented here is the raw data with some missing values.

This resulted in data set, containing 238 storage locations, 117 productin sites, and 5146 pipe segments resulting in 220,000 km of gas pipelines through Europe. Such a data set could be used for static gas flow models, and it is hoped that this data set will be used by the modelling community for answering some of the questions that could arise as part of the Energiewende, and other real world problems.

CHAPTER

FIVE

APPENDIX

5.1 Glossary

Dataset abreviations can be found in Table 5.1.

Table 5.1: Dataset abreviations

Name	Abbreviation	Description
Raw InternetDaten data set	INET	This is the label/name for the raw InternetDaten data set
Raw Gas Infrastructure Europe data set	GIE	This is the label/name for the raw Gas Infrastructure Europe data set
Raw Gas Storage Europe data set	GSE	This is the label/name of the raw Gas Storage Europe data set
Raw Norwegian data set	NO	This is the label/name for the raw Norwegian data set
Raw Long-term planning and short-term optimization data set	LKD	This is the label/name for the raw Long-term planning and short-term optimization data set
Raw International Gas Union data set	IGU	This is the label/name for the raw International Gas Union data set
Raw EntsoG-Map data set	EMAP	This is the label/name for the raw EntsoG-Map data set
Merged and filled IGG data set	IGG	This is the filled data sets, for which the **I**NET, **G**IE and **G**SE data sets were merged
Merged and filled IGGI data set	IGGI	This is the filled data sets, for which the **I**NET, **G**IE, **G**SE and **I**GU data sets were merged
Merged and filled IGGIG data set	IGGIG	This is the filled data sets, for which the **I**NET, **G**IE, **G**SE, **I**GU and the **G**B data sets were merged

The glossary terms can be found in Table 5.2.

Table 5.2: Glossary (A)

Name	Abbreviation	Description
component		A gas network consists of different components, such as pipelines, compressors LNG terminals and more. However, for a gas transmission network, there is a handful of components only: pipeline, compressor, LNG terminal, storage, entry point, border point, connection point, consumer, node, and production
element		Elements are instances of component. Hence, we speak of 10 compressor elements, if we have a data set that has 10 compressors. Here then we can refer to the first or the last or any element of such component
attribute		Gas facilities, such as pipelines or compressors, can be described with a large number of parameters, such as pipeline diameter, or compressor capacity. Those parameters are referred to as attributes. Hence, each component has a list of properties, which are different from one component to another component
facility		General term used for a gas appliance, such as compressor element, or LNG terminal
PipeLine		This is a gas pipeline entity, which has one start and one end point, however can run via many nodes, compressors and other gas network elements
PipeSegment		This is a gas pipeline, that has only one start and one end point, but no nodes in-between, Hence, only goes from one node to another node
LNG	LNG	Liquefied natural gas
CNG	CNG	Compressed natural gas
flow duration curve	FDC	It is the cumulative frequency curve that shows the percent of time specified flow were equal or exceeded during a given period. The information on occurrence of events is lost
Energiewende		German term for the change in using primary energies, the move away from coal to renewable energies, such as wind or solar
gas component data set		Raw input data, associated with components of the gas transmission grid
gas network data set		Output data, a coherent network of gas transmission components
OSM	OSM	Data that is available from the openstreetmap.org
non-OSM	Non-OSM	Data that is not part of the OSM data set
gas type		There are two types of gas High (H) and Low (L) calorific gas
mean absolute error	MAE	mean difference between input values and estimated values
data density		This is the ratio of the number of usable (not missing) attribute values over number elements of the component, in units of [%]
Transmission System Operators	TSO	This is an entity entrusted with the transportation of natural gas/electricity, as defined by the European Union
gas transmission network		This describes the physical gas transmission grid, however excludes any facilities/components that would be part of a distribution network and their facilities. This projects goal is to create an open source gas network data set that can be used to describe the European gas transmission network

Table 5.3: Glossary (B)

Name	Abbreviation	Description
gas component data set		The term "gas component data set" is used for all raw data of objects/facilities that have been loaded using SciGRID_gas tools into a Python environment. However, not all elements (e.g. compressors) must be connected to pipelines, Hence, such a data set is referred to as a "gas component data set", and the emphasis is on the term **component**
gas network data set		A "gas component data set" can be converted into a "gas network data set", by connecting all non-pipeline elements to nodes and all nodes are connected to pipelines, and as part of the process all network islands have been removed, resulting in a single network. Therefore the network contains nodes and edges which are connected, and all objects with the exception of pipelines are associated with nodes in this network, whereas pipelines are associated with edges. Hence, the emphasis here is on the term **network**

5.2 Unit conversions

Table 5.4: Unit conversions

From Unit	To Unit	MultiVal
LNG Mt	LNG Mm^3	2.47
gas tm^3/h	gas Mm^3/d	24/1000
LNG Mm^3	gas Mm^3	584
LNG t	gas Mm^3	1442.48

5.3 Location name alterations

Location names should be changed into the 26 letters used in the English language.

For names from the individual countries please follow the suggested approach:

- Germany/Austria: *Umlaute* to be replaced with the letter followed by an 'e', e.g.: ü = ue.
- France/Belgium: Omit accent de gues and accent de graphs, e.g.: ó = o.
- Sweden: Please change the last three letters of the Swedish alphabet and replace e.g.: ä = a.
- Poland: Please change any letter, that cannot be found in the English alphabet, knowing that for some letters, that one can only use a single letter instead of the three different letters used in the Polish alphabet, e.g.: z = z.
- Spain/Portugal: Please change any letter, that cannot be found in the English alphabet, e.g.: ñ = n.
- Greece: Please do not use Greek letters. Please try to write the Greek words with Latin letters.
- Denmark: Please change any letter that contains non-English letters, e.g.: "å" with "aa".
- Slovakia, Czech Republic, Hungary, Rumania, Latvia, Lithuania, Estonia, Bulgaria, Slovenia, Croatia: PLEASE use your common sense, based on the examples from the other countries above.

5.4 Country name abbreviations

For convenience we provide a short list of names and two-digit codes (see Table 5.5) for the probably most important countries associated with the European Transmission Grid.

Table 5.5: Country codes

Country name	Country code	Country name	Country code
Albania	AL	Kosovo	XK
Armenia	AM	Latvia	LV
Austria	AT	Liechtenstein	LI
Azerbaijan	AZ	Lithuania	LT
Belarus	BY	Luxembourg	LU
Belgium	BE	Malta	MT
Bosnia and Herzegovina	BA	Moldova	MD
Bulgaria	BG	Montenegro	ME
Croatia	HR	Netherlands	NL
Cyprus	CY	Norway	NO
Czech	CZ	Poland	PL
Denmark	DK	Portugal	PT
Estonia	EE	Romania	RO
Finland	FI	Serbia	RS
France	FR	Slovakia	SK
Georgia	GE	Slovenia	SI
Germany	DE	Spain	ES
Greece	EL	Sweden	SE
Hungary	HU	Switzerland	CH
Iceland	IS	Turkey	TR
Ireland and Northern Ireland	IE	Belarus	UA
Italy	IT	Great Britain	GB
Russia Federation	RU	Europe	EU

5.5 Acknowledgement

We acknowledge the contribution of Dr. Ontje Luensdorf from the DLR Institute of Networked Energy System to the SciGRID_gas project.

BIBLIOGRAPHY

[AFW14] M. Ahmed, B.T. Fasy, and C. Wenk. *New Techniques in Road Network Comparison*. Penguin Random House, New York, NY, 2014.

[AG99] H. Alt and L. Guibas. *Discrete geometric shapes: matching, interpolation, and approximation-a survey*. Sack JR, Urrutia J, Handbook of Computational Geometry, Elsevier, New York, NY, 1999.

[DPM20a] J.C. Diettrich, A. Pluta, and W. Medjroub. *SciGRID_gas: The INET gas transmission network data set*. DLR-Institut für Vernetzte Energiesysteme e.V., Oldenburg, Germany, 2020.

[DPM20b] J.C. Diettrich, A. Pluta, and W. Medjroub. *SciGRID_gas: The combined IGG gas transmission network data set*. DLR-Institut für Vernetzte Energiesysteme e.V., Oldenburg, Germany, 2020.

[FMWP+17] Kunz F., Kendziorski M., Schill W.-P., Weibezahn J., Zepter J., von Hirschhausen C., and Hauser P. *Electricity, Heat, and Gas Sector Data for Modeling the German System*. Deutsches Institut für Wirtschaftsforschung, Daten Documentation 92, Berlin, 2017.

[GH85] R.L. Graham and P. Hell. On the history of the minimum spanning tree problem. *Annauls of the History of Computer*, 7(1):43–57, 1985. doi:{10.1145/2729977}.

[Hel18] D. Helle. OpenStreetMap - Deutschland. https://www.openstreetmap.de/, 2018. Accessed: 2019-12-12.

[Kha13] Y. Khalid. What is Pickle in python? https://pythontips.com/2013/08/02/what-is-pickle-in-python/, 2013. Accessed: 2019-10-10.

[LSS+19] P. Lustenberger, F. Schumacher, M. Spada, P. Burgherr, and B. Stojadinovic. Assessing the performance of the european natural gas network for selected supply disruption scenarios using open-source information. *Energies*, 12(4685):1–28, 2019. doi:{10.3390/en12244685}.

[MMK16] C. Matke, W. Medjroubi, and D. Kleinhans. SciGRID - An Open Source Reference Model for the European Transmission Network (v0.2). https://power.scigrid.de, 2016. Accessed: 2019-09-09.

[San19] B. Sandvik. World Borders. http://thematicmapping.org/downloads/world_borders.php, 2019. Accessed: 2019-07-07.

[SAB+17] M. Schmidt, D. Aßmann, R. Burlacu, J. Humpola, I. Joormann, N. Kanelakis, T. Koch, D. Oucherif, M.E. Pfetsch, L. Schewe, R. Schwarz, and M. Sirvent. *GasLib—A Library of Gas Network Instances*. 2017. doi:{10.3390/data2040040}.

[sl19] scikit-learn. 1.1. Linear Models (scikit learn). https://scikit-learn.org/stable/modules/linear_model.html, 2019. Accessed: 2019-08-08.

[UoO14] USA University of Oregon. Comparing distributions: Z Test. http://homework.uoregon.edu/pub/class/es202/ztest.html, 2014. Accessed: 2020-07-07.

[Wik20a] Wikipedia. Bootstrapping (statistics). https://en.wikipedia.org/wiki/Bootstrapping_(statistics), 2020. Accessed: 2019-06-06.

[Wik20b] Wikipedia. Cross-validation (statistics). https://en.wikipedia.org/wiki/Cross-validation_(statistics)#Exhaustive_cross-validation, 2020. Accessed: 2019-07-07.

[Wik20c] Wikipedia. Jackknife resampling. https://en.wikipedia.org/wiki/Jackknife_resampling, 2020. Accessed: 2019-08-08.

[Wik20d] Wikipedia. Lasso (statistics). https://en.wikipedia.org/wiki/Lasso_(statistics), 2020. Accessed: 2020-04-04.

[Wik20e] Wikipedia. Limited-memory BFGS. https://en.wikipedia.org/wiki/Limited-memory_BFGS, 2020. Accessed: 2020-06-06.

[Wik20f] Wikipedia. Out-of-bag error. https://en.wikipedia.org/wiki/Out-of-bag_error, 2020. Accessed: 2019-07-07.

[Wik20g] Wikipedia. Transmission system operator. https://en.wikipedia.org/wiki/Transmission_system_operator/, 2020. Accessed: 2019-09-09.

[Wik20h] Wikipedia. JAGAL. https://en.wikipedia.org/wiki/JAGAL, 2020. Accessed: 2020-01-01.

[YEZ20] I. Yueksel-Erguen and J. Zittel. Approach of converting a PDF map into shape file. priv. comms, 2020.

[BMWi11] BMWi. Forschung für eine umweltschonende, zuverlässige und bezahlbare Energieversorgung. https://www.bmwi.de/Redaktion/DE/Publikationen/Energie/6-energieforschungsprogramm-der-bundesregierung.pdf?__blob=publicationFile&v=12, 2011. Accessed: 2019-02-02.

[BMWi20] BMWi. Home page of BMWi. https://www.bmwi.de/Navigation/DE/Home/home.html, 2020. Accessed: 2020-03-03.

[BundesregierungDeutschland20] Bundesregierung Deutschland. Home page of Bundesregierung Deutschland. https://www.bundesregierung.de/Webs/Breg/DE/Themen/Energiewende/_node.html, 2020. Accessed: 2020-01-01.

[EntsoG20] EntsoG. Home page of EntsoG. https://www.entsog.eu/, 2020. Accessed: 2020-03-03.

[GasIEurop20] Gas Infrastructure Europ. Home page of Gas Infrastructure Europ. https://agsi.gie.eu, 2020. Accessed: 2020-01-01.

[Gassco20a] Gassco. Data page of facilities from Gassco. https://www.npd.no/en/about-us/information-services/available-data/map-services/, 2020. Accessed: 2020-01-01.

[Gassco20b] Gassco. Home page of Gassco Norway. https://www.gassco.no/en/, 2020. Accessed: 2020-01-01.

[IGU20] IGU. Home page of International Gas Union. https://www.igu.org/, 2020. Accessed: 2018-10-01.

[nationalGrid20] nationalGrid. Home page of National Grid UK. https://www.nationalgrid.com/uk/, 2020. Accessed: 2018-10-01.