Improving the topology of an electric network model based on Open Data

Bart Wiegmans
2016

Master Programme Energy and
Environmental Sciences, University of Groningen

university of groningen / faculty of mathematics and natural sciences / energy and sustainability research institute groningen

Research report of Bart Wiegmans

Report: EES-year-number provided by secretary

Supervised by:
Dr. C.B. Davis, IVEM
Prof. dr. R.M.J Benders, IVEM

University of Groningen

Energy and Sustainability Research Institute Groningen, ESRIG

Nijenborgh 4
9747 AG Groningen
T: 050 - 363 4760
W: www.rug.nl/fwn/research/esrig

ACKNOWLEDGEMENTS

I would like to thank Chris for the many interesting and informative discussions, technical and intellectual support, and for leaving me with considerable freedom to research this topic. I would also like to thank Next Energy, in particular David, Wided and Carsten, for allowing me to pursue this topic in a conductive and friendly environment.

Table of Contents

SUMMARY

This thesis report investigates the possibility to construct a network model of the European electricity grid from the OpenStreetMap online geographic database. Network models are essential tools to study the performance of the power grid in energy research. It is based on the work done on the SciGRID model developed at the Next Energy research institute in Oldenburg, Germany. Because the availability of data required by the SciGRID abstraction process in OpenStreetMap is limited, the SciGRID model cannot be extended beyond Germany. The method developed for this thesis relies only on spatial and electrical features of individual objects in the OpenStreetMap. Such features are universally available for objects representing the power system in OpenStreetMap.

Applied to the German high-voltage network this method provides good agreement with human-authored connections used in SciGRID. The network generated can find equivalent or better paths between any two stations in either system in over 90% of cases. This implies that the network is more topologically complete than the network abstracted by SciGRID. However, it is less electrically complete, including more lines and stations with missing electrical information than does SciGRID. It is also considerably more complex, with over twice the amount of stations, more than 30 times as many line joints, and 4 times as many lines as in SciGRID.

The method developed in this research can be applied to the wider European area. This enables open research on the role of the power grid in the European energy transition. However, for application in practical power flow studies more information is required, particularly on power demand and supply in the network. It also yields results which can be used as feedback for OpenStreetMap, for instance highlighting power lines which are not connected to any station or stations which are not connected to any lines.

SAMENVATTING

In dit verslag wordt onderzocht of het mogelijk is een model van het Europese elektriciteitsnet te construeren uit de openbare geografische database OpenStreetMap. Het modelleren van elektriciteitsnet is een essentieel onderdeel van energieonderzoek omdat het elektriciteitsnet een centrale rol speelt in de energie-infrastructuur en in de energietransitie. Het kader van dit onderzoek is het SciGRID model van het Duitse hoogspanningsnet dat is ontwikkeld aan het Next Energy onderzoeksinstituut in Oldenburg, Duitsland. Omdat de specifieke informatie uit OpenStreetMap die gebruikt wordt door het SciGRID model alleen beschikbaar is in Duitsland is het niet mogelijk om SciGRID toe te passen op het geïntegreerde Europese netwerk. De methodes die voor dit onderzoek zijn ontwikkeld zijn slechts afhankelijk van ruimtelijke en elektrische eigenschappen van OpenStreetMap objecten. Deze eigenschappen zijn universeel beschikbaar, wat deze methode in principe wereldwijd toepasbaar maakt.

Toepassing van deze methode op het Duitse hoogspanningsnet resulteert in een netwerk dat in hoge mate overeenkomt met het netwerk gevonden door SciGRID. In dit netwerk bestaat een equivalent of beter pad tussen 90% van alle combinaties van stations, wat impliceert dat het netwerk meer compleet is dan het SciGRID netwerk. Daar staat tegenover dat wat betreft elektrische eigenschappen de informatie in dit netwerk relatief minder compleet is. Bovendien is het veel complexer, met meer dan tweemaal het aantal stations, meer dan 30 keer meer lijn-op-lijn verbindingen, en 4 maal zoveel lijnen als in SciGRID.

Deze methode is toepasbaar op het Europese netwerk als geheel. Dit maakt het mogelijk om open onderzoek te doen naar de rol van het elektriciteitsnet in de Europese energietransitie. Voor praktisch onderzoek aan krachtstroom in het netwerk is er echter nog meer informatie noodzakelijk betreffende de productie en gebruik van stroom in het netwerk. Het is ook mogelijk om de verkregen resultaten te gebruiken als feedback voor het OpenStreetMap project, waarmee de kwaliteit van deze database kan worden verbeterd.

LIST OF TERMS AND ABBREVIATIONS

Algorithm: Contrary to public perception, an algorithm not a magic black box whereby chaos is transformed into money. It is rather a specific procedure consisting of primitive operations that transform inputs into outputs.

Buffer: The area that lies within a specified range from a geometric shape. For example, the buffer of one meter around a point is a circle with radius of one meter centered on the point. Buffers are used to allow inexact spatial matches.

Conductor bundle: A single isolated bundle of conducting elements from a power line.

Subconductor: A single electricity-conducting element in a conductor bundle.

Edge*:* An abstract 'connection' between exactly two nodes, commonly represented as a *pair* of nodes. See also *graph* and *node*.

Geometry: A generic name for various geometric forms and shapes, including shapes that consist of a collection of different forms. The simplest geometry is a point; polygons are examples of more complex geometries.

Georeference: Assignment of a geographic location to an object of interest, in this case, a power station.

Graph: A general-purpose abstraction for 'things that are connected to other things'; similar to 'network'. A graph is formally defined as a pair consisting of a set of *nodes* and *edges*, i.e. G = {N,E}. See also *node* and *edge.* A graph is called *directed* when there is a direction for each pair (e.g. one-way streets), *undirected* otherwise. All graphs described in this report are undirected.

Heuristic: A 'rule of thumb' that allows a decision to be made about ambiguous information. For example, assuming that power lines which end at a substation are connected to that substation may be reasonable, but not a certain fact.

Joint: A place where three or more power lines are connected together, which is not already identified as a power station. Short for 'join-point'.

Node: A node is the name for any single 'thing' which is connected in a 'graph'. Nodes are connected to one another by *edges*. In the context of OpenStreetMap, a node means a 'point on the map'.

Substation: An installation where electrical power is either being delivered to or withdrawn from the electricity grid, including the high-voltage transmission grid.

VGI: Volunteered geographic information. A term for geographic information collected by volunteers.

OpenStreetMap: An online project of volunteers which aims to provide publicly available mapping information.

OSM: Abbreviation of OpenStreetMap.

Path: A sequence of edges in a graph which connect two nodes.

Pseudocode: A simplified representation of an algorithm which serves to illustrate it's general workings rather than the specific details of a particular programming language.

Relation: A collection of OpenStreetMap objects that are organized into ‘roles’. This allows OpenStreetMap to contain complex composite objects such as national borders or bus routes.

Terminal: The endpoint of a line. Lines have two terminals, one of which is designated the 'start' terminal and the other is designated the 'end' terminal.

TSO: Transmission System Operator. A company or organization that manages the electric power transmission infrastructure in a particular area.

Way: OpenStreetMap name for a sequence of points that form either a line or a polygon.

1. INTRODUCTION

1.1 The Importance of power grid network models

The introduction of large-scale renewable energy generation into the electric power systems of the world comes with many challenges. On a global scale, solar and wind energy are abundantly available. However their supply is intermittent and dependent on local (meteorological) conditions. Also, the power that can be captured on any given area is limited. Thus, in order to supply the energy requirements of society and to do so reliably, renewable energy installations must be geographically dispersed. The electric power transmission grid is necessary to balance supply and demand between distributed geographic areas. Before the renewable energy transition can be implemented, it is therefore necessary to determine if the electrical grid can perform this balancing role. In other words, the question is if Spanish sunlight can compensate for a wind-free day in the Netherlands, as well as the other way around.

Because of this distributed nature, many diverse actors are involved in the transition to renewable energy. These actors treat the power grid as a common infrastructure, and they should be able to make informed decisions based on it. For instance, policymakers may need to decide on the optimal location for the placement of a wind turbine park. Private investors and energy companies will want to know that the power generated at such a park can be delivered to customers. Energy researchers are interested in the effects of a particular energy scenario on system reliability. Transmission system operators need to know the effect of shutting down a power line for maintenance. Computer simulation of power flows is an essential instrument in answering these diverse questions.

1.1.1 Basic structure of the power grid

The power grid is at its core a highly complex, highly parallel electric circuit, operating at alternating current. The physical structures of the power grid are diverse and include generators, transformers, overhead power lines and substations. Different parts of the system operate at different voltage levels. Overhead power lines typically operate at a high voltage level, allowing small currents to transmit large amounts of power. These overhead power lines and the structures they connect are referred to as the transmission system, to contrast it with the distribution system. The endpoints of the transmission system are the substations and power plants that consume and produce electric power.

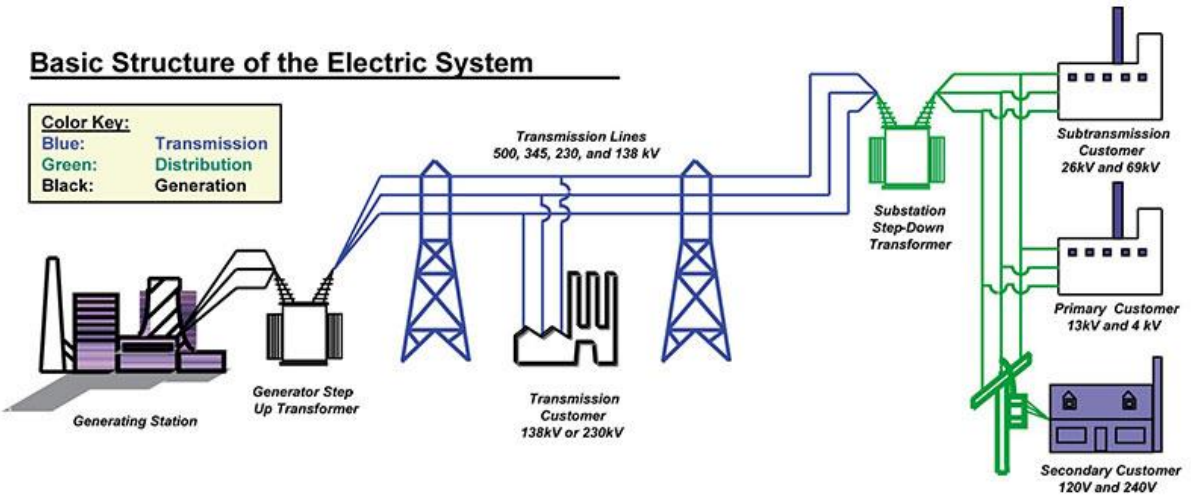

Figure 1 - Structure of the power system. Taken from
http://www.webpages.uidaho.edu/sustainability/chapters/ch06/ch06-p3a.asp

Contrary to Figure 1, a typical power grid contains a multitude of power generators and transmission lines. This makes the grid reliable by providing redundant ways for power to flow and resilience against the failure of individual power plants. The connections between the power plants and the substations and the network that they form allow power generators to 'share' the load of all electricity consumers.

However, the amount of electric power that can flow over a single line is limited. This is because a power line acts as a resistor, although with low resistance. When a current flows through a conductor, power is lost due to the resistance, which heats the power line. As a power line becomes warmer, its resistance increases. This leads to increasingly greater power losses, until the line can no longer efficiently transmit power. Power lines thus have a *thermal limit*, which limits the amount of power that can be transmitted over them. Moreover, all elements connected to the power system must remain in synchrony at all times, which further limits the ways in which power can flow. Computers can simulate the power grid and can determine whenever the limits of the power system will be overrun.

1.1.2 Power network simulations

Power flow simulations require a network model as input. A network model is in its basic form a list of 'nodes' or power stations along with the lines that connect them. Along with information on power supply and demand the flows over the network can be computed and optimized for different scenarios. This is essentially similar to solving circuit equations (using Kirchhoff's laws), albeit much more complex due to the use of alternating current. Software to compute and optimize power flows is readily available online. However, suitable network models are currently not publicly available.

The transmission operators of Europe are organized in the ENTSO-E (European Network of Transmission System Operators of Electricity). ENTSO-E maintains highly detailed transmission network models which are used privately and shared between TSO's (Semerov et al, 2015). ENTSO-E does not share these models with the public over concerns of confidentiality by its members. Researchers may apply to ENTSO-E to acquire a copy. However, this requires signing a non-disclosure agreement, which prohibits the sharing of methodology and open discussion of results. This severely limits the value of such models for scientific research.

An open model of the European transmission system has been made available in 2005 (Bialek and Zhou, 2005) and was updated in 2009 (Bialek and Hutcheon, 2009). This model was created by tracing lines on the map of the transmission system published by ENTSO-E. Unfortunately the network model does not have accurate geographical information. This is a problem for researchers who would like to model for example the effect of specific weather conditions on the stability of the power grid in renewable energy scenario's. Another disadvantage is that this model is quickly outdated by new developments in the power system. Other publicly available network models are sometimes available, but only for a limited area, for example the network model of individual TSO's. (OpENMod Wiki: Transmission network datasets, 2015). Because the European power system is highly integrated, such geographically limited network models are also of limited value. Thus, none of the available network models are suitable for research on the integration of renewable energy in Europe.

1.2 The opportunity of Volunteered Geographic Information

In recent years, the OpenStreetMap project has collected a wealth of information about all aspects of our environment. The infrastructure of the electric power system, consisting of towers, overhead power lines, power plants and substations form an integral part of that environment. As a result, in many areas the physical infrastructure of the power system has been added to OpenStreetMap. The OpenStreetMap database can be freely downloaded in full and there are many tools available for

processing and analyzing it. In Europe alone, over 300.000 separate power lines have been added to the map. This makes OpenStreetMap a valuable resource of power system information for research.

There are nevertheless many challenges involved with using OpenStreetMap data. OpenStreetMap is an open project whereby there is no single authority deciding what goes in and how features should be recorded. The same physical structures may be recorded in various different ways depending on the ideas and preferences of individual contributors. The OpenStreetMap community works together to ensure that the database meets certain standards. However, even then there is a lot of variety in the dataset.

The freely available information on the power system presents an opportunity to develop a network model from this data. Power information in OpenStreetMap typically describes individual structures (e.g. overhead power lines and nuclear power plants). To construct a network model it is necessary to know how these structures are connected. For this purpose specific 'power route' objects have been introduced into OpenStreetMap that record the connections between power lines and stations. In 2014 the SciGRID project was started to take advantage of this opportunity and produced a first network model of Germany in 2015 (Medjroubi and Matke, 2015a). However, the power route objects have not yet become an accepted OpenStreetMap standard. In fact, such objects have scarcely been introduced outside of Germany (Medjroubi and Matke, 2015b). This suggests that the introduction of power routes may be specific to the German OpenStreetMap community. This limits the utility of SciGRID to Germany as well, far short of its ambition to provide a network of the entire European power grid.

However, it may not be unreasonable to assume that when power lines and stations overlap that this implies a connection. This relies on the idea that OpenStreetMap contributors are accurate as well as the idea that the accidental spatial overlap between a power station – especially larger substations and power plants – and (high-voltage) power lines is rare in reality. If this assumption (heuristic) is reasonable then by the full set of these spatial connections it should be possible to extract an electrical network model from the OpenStreetMap database without relying on human-authored connections. The spatial and electrical features of objects that are used are present by definition, for otherwise they would not be part of the map. Due to the variety of ways in which power system structures may be mapped, a significant amount of processing is required before each power line and station is simple enough to form a single edge or node in a well-formed network (Spatialite Wiki: Graph Intro, 2014).

Because the information on which this network is based is provided by volunteers, it is reasonable to question whether it is sufficiently reliable to be a scientific instrument. This is a difficult question to answer without actually using it as a scientific instrument; that is to say, as the input network to a power flow study. Such studies are complex to perform, require extensive data outside the network model itself, and are sensitive to errors. It is therefore a more practical question to ask whether the topology derived by such a heuristic method is complete. In order to answer that, it is only necessary to compare it to another model of the same network. SciGRID provides an acceptable reference network since it is based on completely different method of abstraction using human-authored network paths. In other words, comparing the heuristic method against SciGRID is comparing humans against computers on the same dataset.

Comparing two different networks is not straightforward. Because the methods to derive them are distinctly different, the networks themselves differ substantially. It is too strict a condition to expect perfect line and station equality. At the same time, it is not enough to simply count stations and sum line lengths, for that does not imply functional network equivalence. Such 'functional equivalence' may be found in comparing paths. Two networks which describe the same things at different levels of resolution would be expected to have practically equivalent paths between the same nodes. If paths are not equivalent, this may mean that edges are missing in one network and present in

another. The similarity of paths in two networks may be a measure of the degree to which both networks are equivalent and complete.

In summary, the research community and wider public require a power system network model for the study and management of the renewable energy transition. Such a model is currently not available. Volunteered geographic information in the form of the OpenStreetMap database provides an opportunity to create such a model.

1.3 Research Aim and Questions

The aim of this research is to determine if heuristic abstraction of the power network can be used to improve the topology of the SciGRID network model and extend its application to the continental area of Europe. To achieve this, this report answers the following three research questions:

1. How can power lines and stations mapped in OpenStreetMap be processed to form a well-formed electrical network?
2. How can this network be compared to other networks, especially to SciGRID?
3. Is it possible to apply this method to a wider area, especially to Europe?

2. METHODS

2.1 Process Overview

The process described in this section derives the topology of an electrical network from the map objects provided by OpenStreetMap. In order to distinguish it from the SciGRID method, it has been named 'GridKit'. Like SciGRID, GridKit is an open source project which can be downloaded from the internet (GridKit, 2016).

The OpenStreetMap database is primarily used for the purpose of making maps. Deriving a network model from map data is ultimately a matter of processing map objects to form simplified lines and stations. When a map is made, typically little concern is given to the internal structure of objects in a map. For example, a single power line may be drawn in multiple parts, which may or may not be connected. Different lines may share a part of their part and then split, and this can be represented in various different ways. Similarly, a single 'logical' power station may be drawn in multiple parts, sometimes in great detail. Correctly interpreting such a variety of structures is next to impossible, hence most of the effort in the derivation process is directed to reducing this variation.

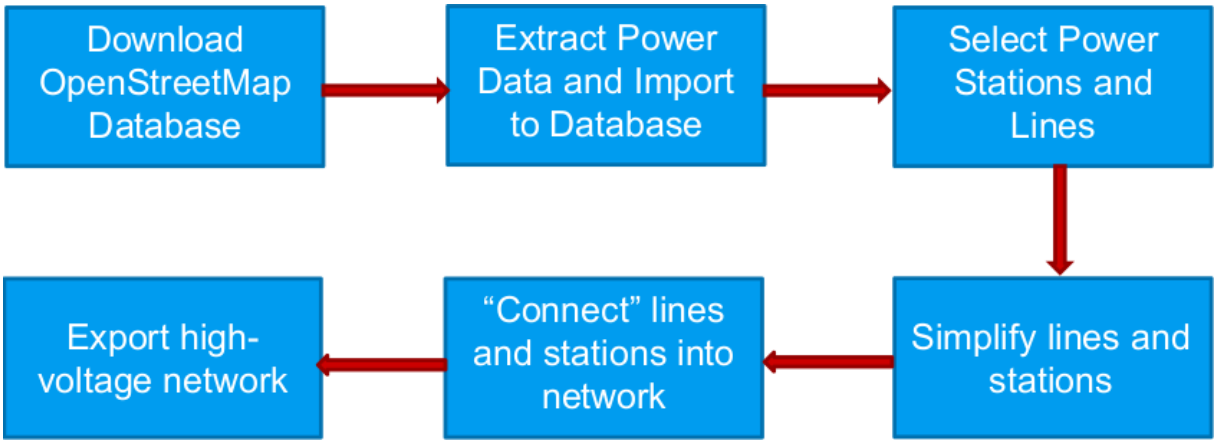

Figure 2 - Overview of GridKit process

The OpenStreetMap database contains a very large number of objects describing power system structures, but it does not in general describe their relations unambiguously. For example, although an overhead power line may be mapped as spatially overlapping with, it is not definite that this implies an electrical connection. To assume so may be reasonable, but it is nevertheless an assumption, and it serves to add information to the resulting network topology that was not present in the source database. The degree to which the results of this procedure are reasonable depends accordingly to the degree to which these assumptions hold in practice. As the OpenStreetMap source database may change, this too may change. A limited number of core assumptions guide the strategy of the derivation procedures:

- Power lines and power stations rarely overlap accidentally. When they do overlap, this will usually imply they are connected.
- This overlap may not be recorded exactly in the OpenStreetMap database, therefore it is reasonable to extend the geographical region of intersection of lines and stations by applying a modest *buffer*.
- Power lines that form a single electrically connected line, may be recorded as multiple lines in OpenStreetMap; these lines may or may not represent physically distinct structures. For example, a power line may be partially overhead and partially underground, and still form a single connected line.

- A power line may overlap more than two stations, for example, when a line connects multiple stations in a series. Such a single 'series-line' is equivalent to the set of station-to-station lines that connect the same path.
- Power lines which cross one another are common and do not imply a connection. However, a line which ends at some point on another line represents a point where those lines join together.
- A shared point where more than two lines terminate also represents a point where lines join together.

To allow the unambiguous abstractions of connections, it is necessary to transform the OpenStreetMap objects to *simple lines* and *simple stations*. This simplicity refers to their spatial features. A simple line terminates on at most two distinct stations, thereby connecting them. It does not terminate on another line, nor does it run across a station in its path. Similarly, a simple station spatially overlaps no other station. Arbitrarily many lines may connect with a single station. If lines would not be spatially simple, it would be impossible to determine if a line that connected (for example) 3 stations connect them in a triangle or in a series.

This spatial simplicity does not imply that they should be equally electrically 'simple'. A power station may consume power at one voltage level and provide power at another (i.e. a transformer), or it may convert between AC and DC (i.e. a rectifier). A power line may carry wires that transport power at distinct frequencies or voltages. Thus, the resulting stations and lines may be *electrically* complex but must be *spatially* simple.

In contrast to SciGRID filtering the high-voltage network from the larger power system is only performed at the last possible moment, after the full topology is already known. This allows for the extraction of stations which have no voltage information themselves but which are attached to high-voltage lines as well as lines without voltage information which connect two high-voltage stations. In this sense it is a more optimistic approach than the others.

The derivation procedure consists of the following high-level steps:

- *Identification* of specific power system objects as power lines and stations.
- *Simplification* of OpenStreetMap objects to simple lines and stations.
- *Abstraction* of connections from their spatial overlap

Prior to the derivation process, the map data is acquired in compressed format from an online source, for example OpenStreetMap full planet copies or Geofabrik geographically limited extracts (Geofabrik, 2016). This file is then filtered to contain only structures of the power network. Finally the filtered data is imported into the *PostgreSQL* open-source database using the *osm2pgsql* utility. Automating this process is trivial. In the procedures described here it is assumed that this import has already happened.

2.2 Source Data Model

In order to describe the processes that are applied to OpenStreetMap objects, it is first necessary to describe the map objects themselves. The OpenStreetMap database structure seems to have been designed for simplicity and generality and contains only three different object types.

The basic geographic object in OpenStreetMap is a *node* (or *point*), which describes a single location on the surface of the earth (OpenStreetMap Wiki: Node, 2015). All other objects are ultimately built from nodes and derive their geographical features from them. A node may describe a variety of physical structures, from mountaintops to transformer boxes or the towers supporting high-voltage power lines.

All structures with more complicated geographic features are recorded as *ways* (or *lines*). A way is an ordered list of nodes. A line may form a closed loop, in which case the way forms a polygon (OpenStreetMap Wiki: Way, 2016). All such polygons are simple because complex polygons with inner areas are not allowed. In OpenStreetMap, ways are used to record structures as varied as roads, buildings, power lines and substations. Nodes may be embedded in multiple ways.

Finally, arbitrary relations between map objects may be recorded in appropriately-named *relation* objects. Relation objects can refer to any other map object (nodes, ways, and other relations) and assign *roles* to each of them. Such objects can be used to record information as varied as bus routes, natural areas, and borders. Like everything else in OpenStreetMap, roles are free-form, although the community attempts to enforce certain standards.

Each OpenStreetMap object may be associated with one or more *tags*. A tag is a textual key and value that describes some feature of the object in question. Tags are vital in identifying what an object represents (e.g. to determine whether something is a road or a river) and to record features of them (e.g. the particular voltage on a power line). Tags are entered completely free of restrictions, although the OpenStreetMap community attempts to establish and enforce certain tagging standards.

When dealing with geographic information a very important choice is which coordinate system to use. Not all coordinate systems can accurately record information of all locations on earth. OpenStreetMap uses the standard WGS84 coordinate system (Wikipedia: World Geodetic System, 2015), which uses degrees of latitude and longitude relative to the equator and the international reference meridian. This system can record any location on earth and is also used by GPS systems. The fact that any consumer GPS system can provide coordinates for a geographic point is a considerable advantage for a volunteer project.

Prior to drawing a map, the geographical coordinates of objects need to be 'reprojected' in a suitable projection system (Wikipedia: Map Projection, 2016). The projection system used by OpenStreetMap online maps is a simplified version of the traditional Mercator projection. The Mercator projection system is suitable for global display, although it causes distortion of distances at the latitudes of European countries (ArcGIS blog: Measuring distances when your map uses a Mercator projection, 2010). This projection system is also used for the analysis described in this report. One of the advantages of this system for analysis and processing is that distances can be computed in meters rather than degrees. Accurate distances may be computed after reprojection to a more suitable system. Another advantage is that geometric calculations (for example, to determine whether a point is contained in a polygon) is considerably simpler in a projected system than when using the true three-dimensional surface of the earth. This is a significant concern due to the large amount of processing required by some procedures.

2.2.1 Power System structures in OpenStreetMap

Electrical power systems consist of a great variety of components and structures. Those structures may be classified as power stations or power lines. A power station is a (geographically) fixed point where power is either produced or consumed. A line is a connection between two stations. In literature on power flow modeling a station is also commonly called a 'bus', and lines are called 'branches'. Power stations may contain transformers for converting between voltages, as well as rectifiers and frequency changing electronics. Power stations that operate at multiple voltages or frequencies may be represented as a single (complex) stations or as multiple distinct stations which happen to share a geographical location. In power system simulations, the latter situation is more common, with the converting electronics (e.g. transformers) represented as distinct 'branches'. This allows accurate modeling of conversion losses, which may be significant. Similarly a single power line may carry wires operating at different voltages, which are commonly modeled as distinct lines for the same reason. In OpenStreetMap, power lines and stations are typically recorded as complex

structures; for example, a single power line may be recorded to carry power at 16.7hz and 50hz in different conductor bundles. (The 16.7hz frequency is used by the German electrical railway power network).

Identification of power system structure relies on the value of the standard 'power' tag. The value of this tag usually indicates what type of system is in use. While the 'power' tag key is standard, a large variety of values (more than 100) are in use. The 10 most common 'power' type objects (excluding nodes) are listed in the table below. Power types which are meaningful for the system topology and which can be classified without ambiguity as lines or stations are identified according to the last column.

Table 1 - Most common 'power' object types in OpenStreetMap

Value	Frequency	Identified as
minor_underground_cable	565	Power line
plant	650	Power station
transformer	951	Not identified
station	1.238	Power station
cable	3.083	Power line
sub_station	7.736	Power station
substation	17.667	Power station
line	22.992	Power line
generator	26.020	Not identified
minor_line	36.819	Power line

Very common elements are also 'tower' and 'pole' (246041 and 433607 each, respectively). This is because overhead power lines are typically recorded as sequences of 'towers' and 'poles'. Objects tagged 'generator' are not included in the analysis, primarily because there are so many of them (including nodes there are 57.009 separate generators included in the German database). Many of these generators represent small-scale systems such as residential photovoltaic installations. Such systems are quite likely to accidentally overlap with power lines, and quite unlikely to be connected directly to them. Therefore, they are not considered further here, although they may be a valuable source of information for other purposes.

Other standard tags describe relevant electrical properties of power structures. The most common tags are listed in the following tables. Most power lines and stations do not have complete electrical information. Further complicating the interpretation of electrical information is that such tags can typically contain more than one value. For example, an overhead power line may carry both 110kV and 220 kV wires, bundled into 3 and 6 cables, and this may be recorded accurately. But that still leaves ambiguous which of these cables carries 110kV and 220kV.

Table 2 - Most common tags on power elements in OpenStreetMap

Power Lines		Power Stations	
wires	18443 (29%)	`frequency`	3977 (9%)
frequency	19405 (31%)	`location`	4816 (11%)
operator	19927 (31%)	`voltage`	6526 (15%)
voltage	34247 (54%)	`substation`	11086 (26%)
cables	34922 (55%)	`operator`	13992 (32%)

2.3 Basic Algorithms

One of the central algorithmic steps of the derivation procedures is finding 'connected sets'. A connected set is a group of items that each share some property with at least one of the other items, but not necessarily all of them. We want to find all items that somehow mutually share this property.

This problem arises, for example, when finding all lines joining at a certain point, connecting lines end-to-end, or merging stations that overlap spatially.

The connected set algorithm treats such problems as a graph problem. In the simplest implementation, every item is a node in the graph, and every pair of items sharing a feature form an edge. The process starts out by assigning to each node it's own set. The minimum-valued node in the set is used as the key of the set. Now, when iterating over all pairs, the sets to which each node belongs is looked up. When these are different, both sets are merged together and assigned the least-value key. When the process has ended, all nodes which are connected by a path are then assigned to the same set.

```
procedure connected_sets(nodes, edges):
	let sets be a mapping from the name of each node in nodes to a set containing only node;
	loop for each pair in edges
		let a be the set belonging to the first node of the pair
		let b be the set belonging to the second node of the pair
		if key of a does not equal key of b then
			let c be the result of merging set a and set b
			assign set c to each node in set a and set b
		end if
	end loop
	return sets
end procedure
```

2.3.1 'Shared Nodes' and Spatial Algorithms

As mentioned in the section on OpenStreetMap structures, two or more ways may share a single node. When two or more lines are connected in this way by a single node, it is quite reasonable to assume that they are connected physically. Using this logic, two lines that each share a node at their *endpoints* probably represent a single line that was drawn in two (or more) parts. Similarly, when more than two lines share an a point, or when two lines share a point that is not the endpoint of one of them, that probably represents a place where lines joint together. Using node identity can thus be used to find lines to merge together and places to insert line joints.

Spatial algorithms rely not on node identity but on spatial overlap of structures. As a result, such algorithms are more robust to the different ways in which structures can be drawn. Using *buffered areas* around stations and line ends ensure that connections may be found even when objects do not overlap exactly. Moreover, spatial algorithms can treat the entire structure as a whole, while shared-node algorithms only take a single node of the larger structure into account. Thus, spatial algorithms can be used for merging overlapping stations and lines, and for removing overlapping segments between lines and stations, which is not possible using nodes.

The 'shared-node' algorithms have two main advantages over spatial algorithms. First of all, a node identity search query is considerably simpler, and hence faster to execute, than an equivalent spatial overlap query. Second of all, because they do not rely on buffered areas, they are not sensitive to 'buffer artifacts', which distort the actual connections, especially in areas that contain a cluster of lines. This situation is very common in distribution lines and surrounding substations. However, the shared-nodes algorithm for inserting joins may create 'beads on a string' artifacts, precisely because they only observe a single node at a time. For the best results both types of algorithms are necessary.

2.3.2 Topological Algorithms

The final class of algorithms form the so-called 'topological' algorithms which employ the (derived) connections between stations and lines. Such algorithms can only be used after connections have been found. Therefore, they cannot serve to prepare and simplify power lines and stations. However,

such algorithms can serve to simplify the network and to propagate incomplete information from neighbors. For instance, a topological algorithm is used to detect the 'beads-on-a-string' artifact as well as to find the full set of high-voltage network lines and stations. This is necessary because only a minority of lines carry full voltage and frequency information.

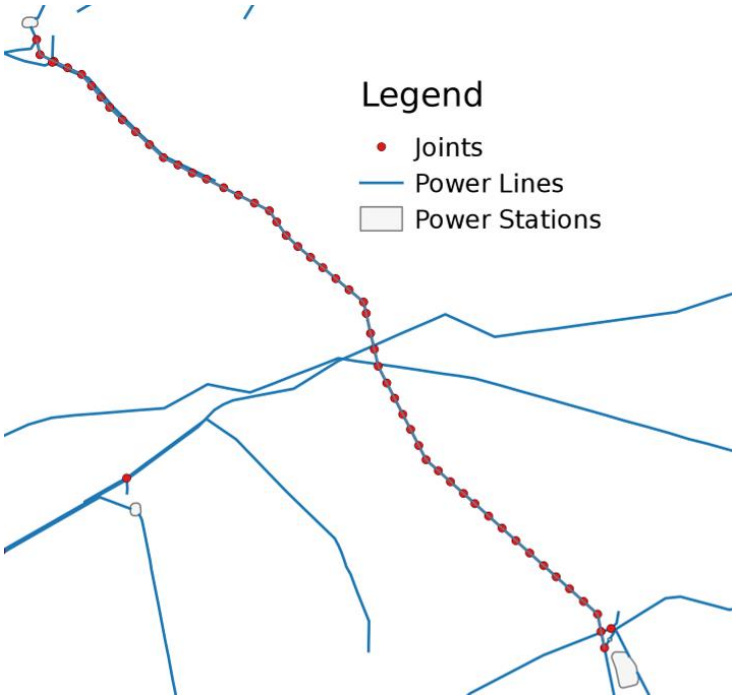

Figure 3 - Beads on a string artifact. This is the result from per-node joint insertion when two lines share a node. The resultant network is not necessarily wrong, but the additional joints add no information. (c) 2015 OpenStreetMap contributors

2.4 Spatial Simplification Procedures

2.4.1 Merging lines and stations

All power stations which overlap spatially with another station are assumed to be part of a single, larger power station. The connected-set algorithm is used for finding the full overlapping set of stations, since station A may overlap with station B and station B may overlap with station C, while station A and C may not.

Lines which connect on each end – except when this connection overlaps with an existing power station – are joined together. See for example Figure 4. Their properties are also merged. This is implemented as a shared-node algorithm as well as a spatial algorithm. Both versions use the connected-set algorithm to find sequences of lines that connect together. The electrical features of the merged line are formed by the merged set of their constituent lines.

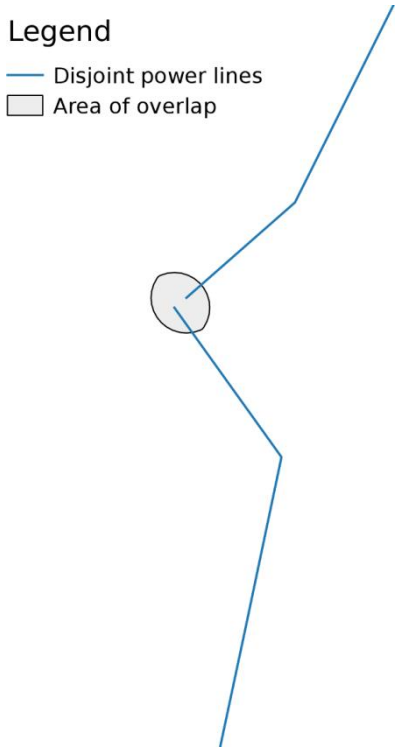

Figure 4 - Merging lines. The area of overlap is formed by the ‘buffers’ applied around the end of each power line, which allows the connection to be determined even when they are not connected directly.

The simplest primitive procedure for connecting lines connects the end terminal of the first line with the start terminal of the second one. When lines are 'misoriented' (meaning that the terminals which are to be connected are not the end and start points of both lines), this introduces a long line segment between the respective end and start points of the lines, artificially inflating the true length of the line. To counter this, a simple heuristic is used whereby both lines are connected in all 4 possible orientations, and the shortest resulting line is chosen as representing the 'connected' line.

2.4.2 Eliminating line-station overlap

In some cases, the power lines which are internal to a power station are drawn in great detail (see Figure 5 for an example). Such details are removed from the topology before further processing. This applies to 'power lines' which lie completely internal to the station (e.g. busbars) as well as the segments of external lines that overlap the station. One of the reasons for this reduction of detail is that the complex internal connections are very likely to upset the spatial connection-finding algorithms. Also, the exact inner details of power substations is usually of little interest. Such structures can be modified during routine operation and maintenance of the substation, which makes it likely that any information abstracted from them are quickly outdated. Further, the correct interpretation of the electrical network is likely to be very difficult if not impossible in spite of the level of detail. Finally, in most cases these internal lines will form a fully connected network, which renders the additional detail redundant.

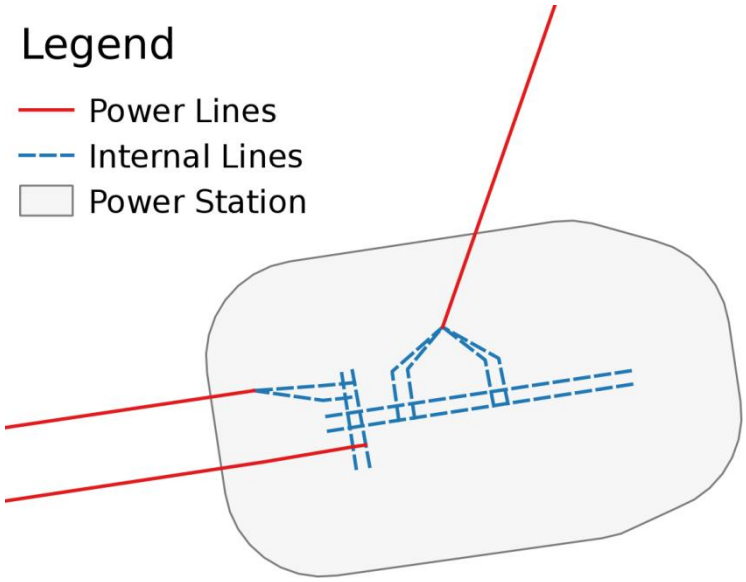

Figure 5 - An example of a highly detailed map of a power station. The internal lines presumably show structures like busbars, but this is difficult to interpret. (c) OpenStreetMap contributors.

In some cases, external lines may cross multiple stations. In this case too, the overlap of the line with each station is eliminated. In this case however the elimination results in a multitude of segments which run between and beyond the stations with which it overlaps. In effect the power line is 'split' into segments between each of the stations it connects. These segments replace the earlier power line, inheriting its electrical features.

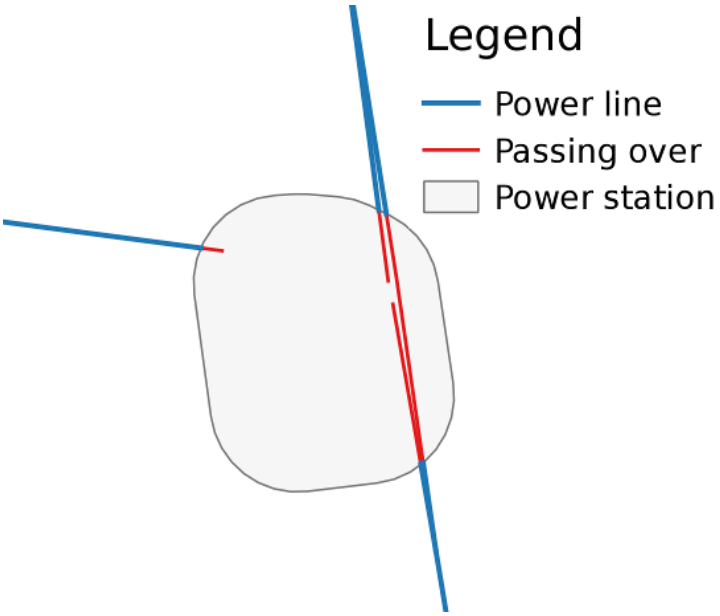

Figure 6 - A line that passes over a station. The internal (red) segment of this line is eliminated, yielding two lines for the north-to-south power line. The ‘split’ just prior to the station is eliminated in a later step.

2.4.3 Inserting joints

Real power lines do not just connect stations but also other lines, for instance at locations where a single power line splits and continues in two directions. Simple lines cannot do this. Hence, wherever more than two lines connect a 'joint' virtual station must be inserted. Note that a power line which is attached 'in the middle' by another line will be counted as two distinct lines, one for each end.

Three distinct algorithms have been implemented for the purpose of inserting joints. One is based on shared nodes, the other on attaching lines that end midway on other stations, and the third is based on spatial overlap between line terminals. All these algorithms proceed in a similar fashion. The first step is to find each point where a 'joint station' should be inserted. The shared-node algorithm looks for all nodes that are shared by more than three lines, or nodes which are shared by two lines and are not the endpoint of at least one line (implying that this line has two segments extending from the node, counting as two lines). Attachment points are found by searching for line terminals that overlap with the extent of another line, but not with its terminals. The final algorithm looks for all sets of line terminals which share space and have more than two members. All algorithms exclude points that lie inside power stations.

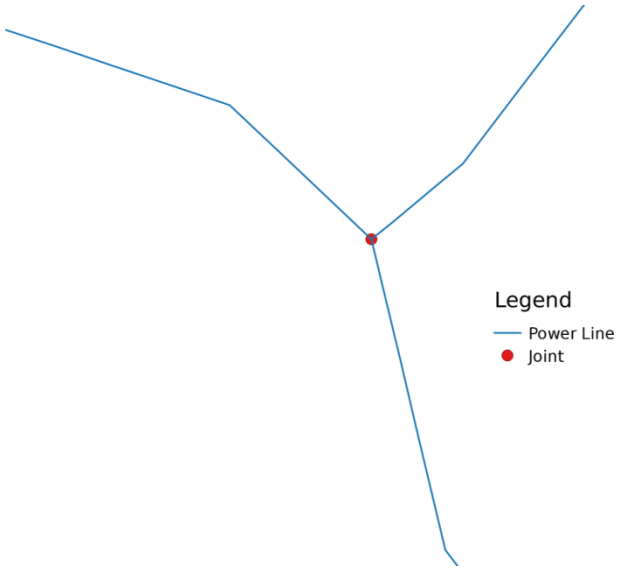

Figure 7 – A joint, in this case inserted on a shared node. The resulting three lines are simple.

When all such joint locations have been found, a virtual station is constructed, by creating a small buffer around the location. The spatial difference between each joining line and the newly constructed power station is then computed. In case this consists of multiple line segments (i.e. when a line is split in the middle by a joint) each new segment is inserted as a new power line, replacing the old line. If a line that forms a joint does not exactly attach to that joint – for example, because of an inexact match by a terminal buffer – that line is extended to connect directly to the joint.

Aside from simplifying lines, joints also serve as 'changing stations', where a power line that carries multiple sets of cables (each of which may carry a different voltage and even frequency) can be split into lines that each carry a set. Many joints are however redundantly inserted, when lines split just before attaching to a power station (see also Figure 8). Such splits do not add to the topological information of the network.

2.4.4 Reducing artifacts

Artifacts are lines, joints, or connections that are introduced to the network a derivation procedure, which do not correspond to map structures, or which do not add information to the network. *Buffer artifacts* are connections which are produced by accident, because the that are applied buffers happen to be larger than the distance between different structures.

The basic strategy for reducing the impact of buffer artifacts is to use buffers as little as possible – for instance, by using shared-node algorithms, which do not employ buffers – and to use buffers that are

as small as possible. When, for instance, a 'joint' station is inserted between lines, it is a certainty that those lines ought to be connected to that particular joint virtual station *and to nothing else*. In this case a minimal buffer of 1 meter radius around the connecting terminal can be computed to decrease the likelihood that an accidental connection to another structure is produced. Similarly, when a line is split along a station, each of the resulting segments is also given a 1-meter terminal buffer connecting to the station. As lines are transformed in further processes, care is taken to preserve the computed terminals, so that false connections are not introduced. For the same reason, a line-end terminal radius is at most 1/3 of the line length in size, to reduce unwanted terminal overlap in shorter lines. The maximum size of a line terminal is 50 meters. Note that the meter units used do not correspond exactly to on-the-ground meters, because of the distortion introduced by the Mercator projection.

Joint artifacts are line joints which are not necessary for the correct rendering of the network topology. Three distinct types of joint artifacts are recognized. The first is the 'redundant joint', which connects only two stations. Such a joint and both its edges can be replaced by a single edge. A special case of a redundant joint is the 'beads-on-a-string' artifact, which results from two or more lines which share a portion of their extent, and which are split by the shared-node joint insertion algorithm. Each shared node in their extent is after all indistinguishable from a node where two separate lines attach in the middle. Note also that the first and last nodes of the beads-on-a-string are typically 'true' joints, as is also visible in Figure 3Figure 3.

The second type of joint artifact recognized are splits-before-stations. Those splits are common when a power station is mapped in great detail, and each segment from the split is drawn as being connected to another internal element. From the perspective of the topology these splits add no information because both edges attach to the same station. Another way to consider this is that if the station area buffer had been larger those splits would have been eliminated as internal details to the power stations, which is essentially what they are.

The final type of joint artifact is a dangling-joint; that is to say, a correctly mapped joint which connects lines which do not themselves connect to stations (i.e. dangling lines). Dangling lines do not partake in the topology, so a dangling joint is a place where lines are split that – from the perspective of the topology – end in nothing. In other words, such joints do not actively participate in the network and can be discarded.

Note that none of these artifacts are really errors, since they correspond to the true situation (or at least to the true map). They are just of very little interest to researchers who would like to study the model, and they don't add topological information. Moreover, these are far from the only types of structures which are of limited value, but rather the structures which can be easily detected. Some edges form small 'triangular' cycles, and these are probably equivalent to a single joint connecting them. Proving that this is possible and reasonable is far more complex, though.

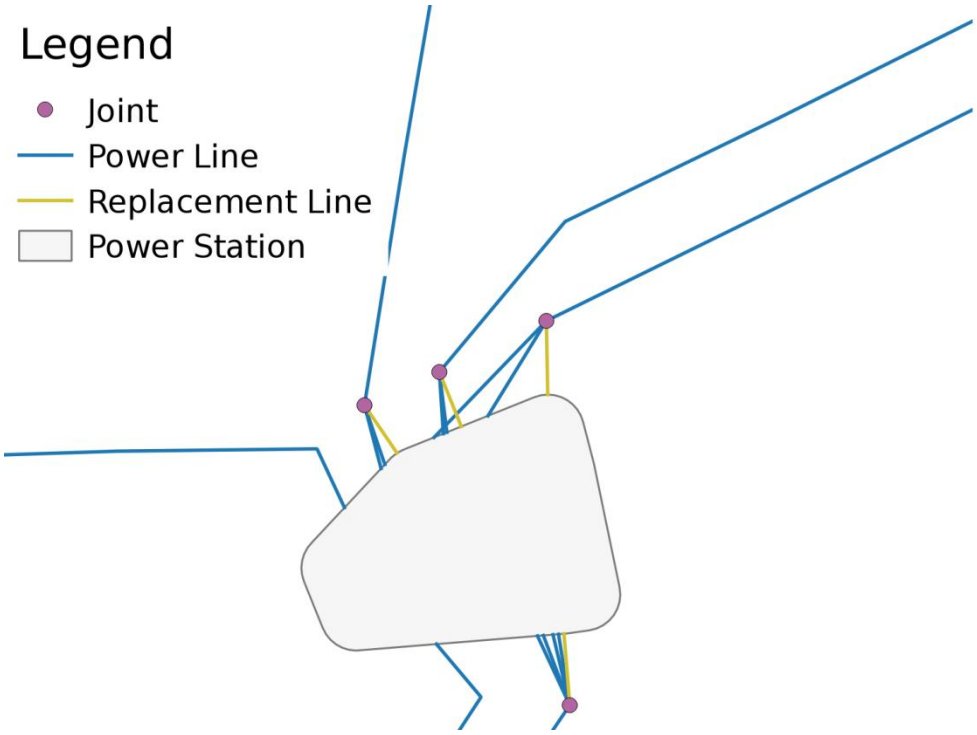

Figure 8 - Simplification of 'split before stations' artifacts. Afterwards, the split connects just two stations, rendering it redundant. These redundant joints are then removed (c) 2015 OpenStreetMap contributors.

Dangling joints can be removed directly. However, after removing joints, new joints may be 'uncovered' which have then become dangling. In this case, those joints can be removed too, up until no more such joints can be found. For this reason it is implemented as an iterative process that continues until no more 'dead' joints can be found. Afterwards, the topological database may contain edges which point to a now-missing joint, and those edges also have to be removed. This in turn may uncover new 'dead' nodes which point to nothing and for that reason can be removed. Fortunately, removing those nodes *cannot* uncover new dangling edges, for that would imply they had to be connected by those edges, and in that case they would not have been removed. For the same reason this cleanup step cannot introduce new dangling joints. So the pruning step is complete.

The 'beads on a string' and 'split-before-station' artifacts each imply that a multitude of lines each pass from the same source station to the same destination station. Those lines are collapsed into a single line if this does not cause a very large distortion of their original lengths (the difference should be less than 300 meters). In the case of the 'beads on a string' this difference will naturally be negligible, but even in 'splits before stations' the distortion is typically very small. Some care has to be taken not to introduce 'double-simplified' lines, when two adjacent 'bead' joints each introduce a single simplified edge to replace the double edge between them. To prevent this, the identifiers of both stations are used together as an identification key for introducing the simplified line.

Finally, redundant joints are removed by treating them as pairs of edges that should be connected in the disjoint-set algorithm. After simplifying the double edges of 'beads on a string', this step also removes the beads and reduces them to a single line that is only split at its beginning and end nodes.

2.5 Extracting the high-voltage network

The topological network derived by GridKit thus far is very large (over 18.000 edges and over 17.000 nodes). Most of this network does not represent the high-voltage lines and stations, which form the primary interest of most researchers. Extracting this high-voltage network has been deferred until the latest possible moment in order to use connection information. This is because many of the lines

and stations in the network contain incomplete information to be included in the high-voltage network. However, if a line connects two high-voltage stations, it is almost certainly operated as a high-voltage line as well. Similarly a station that is connected to a high-voltage line must also operate on a high voltage and thus be included in the network. Only after all the connections have been found can these network elements be included.

During the simplification procedures, lines and stations may be merged together or split into pieces. These new objects acquire their 'tags', and hence their electrical properties, from the original OpenStreetMap objects from which they have been derived. For this purpose a 'tracking table' is implemented. In some cases, the electrical properties of combined lines may seem to conflict, for instance a line that operates at 16.7Hz combined with a line that is operated at 50Hz. While such a combination would at first sight seem to be impossible, such lines have in fact been added to the map, and presumably correspond to the real-world situation. (For instance, a combined power line that runs over a river and splits in different lines afterwards). Thus, it seems to be impossible to automatically detect such 'impossible' combinations and distinguish them from the true situation. This interpretation is deferred to the users of the model.

The final step of the derivation procedure is to convert the complex mapped lines and stations to the simplified nodes and edges that are used by SciGRID. This allows both methods of extraction to be interchanged as well as compared to one another.

2.6 Comparison of Networks

In order to qualify the result of the abstraction process a quantitative method of comparing the constructed networks is required. Because the SciGRID network is constructed from the same source database, but uses human-recorded features rather than implicitly derived connections, it can serve as a reference network. However, the GridKit derivation process constructs a much more complex topology than SciGRID does. This is because SciGRID can create direct edges between each station recorded in a relation, whereas GridKit must create edges between each joint on the way. (SciGRID also inserts joints when they have been recorded in the relation objects, but interpreting these joints is complex and not generalized). Thus in general power lines in one network will not have a direct equivalent in the other. However, power stations will typically derive from the same objects and thus share the same locations. Power stations can thus be identified between network models because they share the same space.

Ideally the two network models should be compared by their performance on an intended application. For these models, that could be various forms of power flow analysis. However, such analysis requires detailed information on power loads and generator capacities, which is not present in the network topology or in OpenStreetMap. The possibility to compute simplified power flows with a single generator and a single consumer station was investigated. However it was found that such simplified flow computations are not solvable with commonly accepted power flow computation methods. Thus, comparing both networks on performance as electrical systems is not feasible without proper real-world electrical information.

However, it is still possible to compare networks on their topological features, given that we are able to correlate stations between both systems. (Without this correlation, this becomes an intractable problem). If both systems accurately capture the German high-voltage power network, it is to be expected that certain statistical features will match closely, for instance the total length of all lines.

More interesting than statistics is the *connectivity* of the network. If both models correctly map the same high-voltage network, it is to be expected that each station connects to roughly the same set of neighbors. Because of the difference in derivation process between SciGRID and GridKit, the exact power lines between stations do not match. In general GridKit construct a more complex network with multiple joints between two stations. Thus it is not very useful to compare the direct neighbors

of each station between both systems. However, the *paths* between stations should be very similar as long as they each correspond to the same power lines. When a particular power line is missing from one network and present in the other this will force the path to take a detour, meaning that it will be significantly longer than the corresponding line. Thus, the relative length of paths can be used as an index for the presence or absence of specific power lines. Naturally, only stations that have an equivalent in both networks are candidates for this comparison. The comparison of shortest paths between stations that are direct neighbors in SciGRID is a special case which serves as a sanity check; in almost all cases, these distances should be very nearly equal, because they follow the same power lines. To find the shortest paths a standard network search algorithm called 'A*' (A-star) is applied.

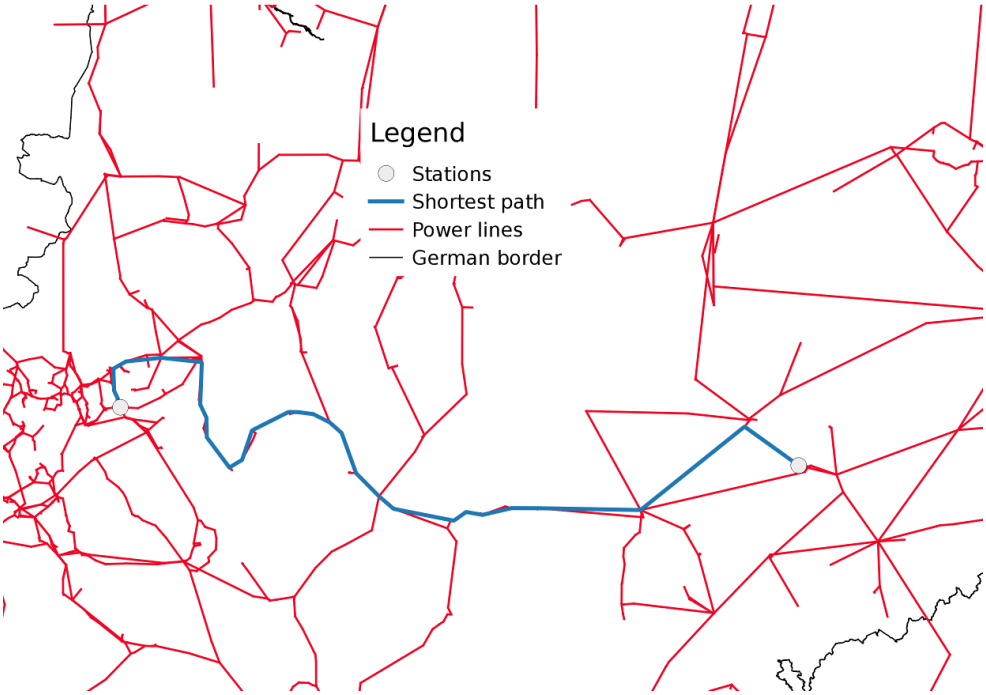

Figure 9 - An example of a 'shortest path' between two stations. Note the ‘detours’ near each station, which may imply a missing connection which is not visible at this resolution.

3. RESULTS

3.1 Network statistics

From the German database used by SciGRID, GridKit extracts 929 high-voltage stations as 1422 high-voltage joints and 3140 high-voltage lines. However, due to a technical artifact from the filtering processed used by SciGRID, the high-voltage network of Poland is fully included in this database. After excluding the Polish network from the GridKit model, 815 stations, 1315 joints and 2810 lines remain, compared to 455 stations, 40 joints and 815 lines. Graphically the GridKit network appears to be very similar to the SciGRID network (see Figure 10).

Of the (in total) 495 nodes in SciGRID, 476 can be identified the GridKit network. Of the 19 unidentified nodes, 16 are generators, which have been explicitly kept out of the analysis. 2 are substations, and 1 is an joint. This implies that 39 of 40 joints in the SciGRID model are placed identically in GridKit. This is remarkable because both models have considerable freedom in placing joints. However, those 476 nodes in SciGRID map to 453 nodes in the GridKit network, indicating that some of them probably map to stations that have been merged together in the derivation process.

The total length of the high-voltage network lines is 30.400 km long, whereas the total length of lines of the SciGRID network is 30.600 km long. This similarity is misleading, however, because the GridKit high-voltage network contains lines spanning the full area of Poland, and the SciGRID network does not. Thus, the high-voltage network should be considerably longer. After filtering the lines to those lying in Germany alone, the network is just 20.300 km long. The full network including all power lines is 83.300 km long, while the full length of all mapped power lines is over 126.000 km, indicating that only 66% of all power lines are part of the well-formed network.

Of the remaining (43.000 km) lines, 34.000 km (79%) is formed by 'dangling' lines, which attach to one station or none. 7600 km (18%) is formed by 'pruned' edges, which attach only to joints. Only 220 km (0.5%) is formed by 'problematic' lines, which even after simplification attach to more than two stations (often because two of these overlap). This leaves 650 km (1.5%) unaccounted for, which is likely the result of lines sharing a portion of their extent, as well as line distortions during simplification.

Table 3 - Statistics of GridKit and SciGRID networks in Germany

	SciGRID	GridKit High-Voltage	GridKit Full network
Number of nodes (joints and stations)	495	2130	17022
Size of largest connected set	490	2111	11619
Number of stations	455	815	5877
Number of joints	40	1315	11145
Number of lines	825	2810	18439
Total length of lines	30.632 km	20.033 km	83.337 km
Average length per line	37.1 km	7.1 km	4.5 km
Voltage (stations)	93% (461)	29% (618)	28% (4781)
Frequency (stations)	69% (340)	20% (431)	17% (4781)
Complete information (stations)	69% (339)	20% (424)	15% (2635)
Voltage (lines)	100% (825)	91% (2564)	77% (14216)
Frequency (lines)	71% (586)	56% (1574)	39% (7111)
Conductor bundles (cables)	96% (800)	92% (2577)	76% (14005)
Subconductors (wires)	79% (655)	74% (2090)	52% (9608)
Complete information (lines)	79% (649)	72% (2032)	50% (5317)

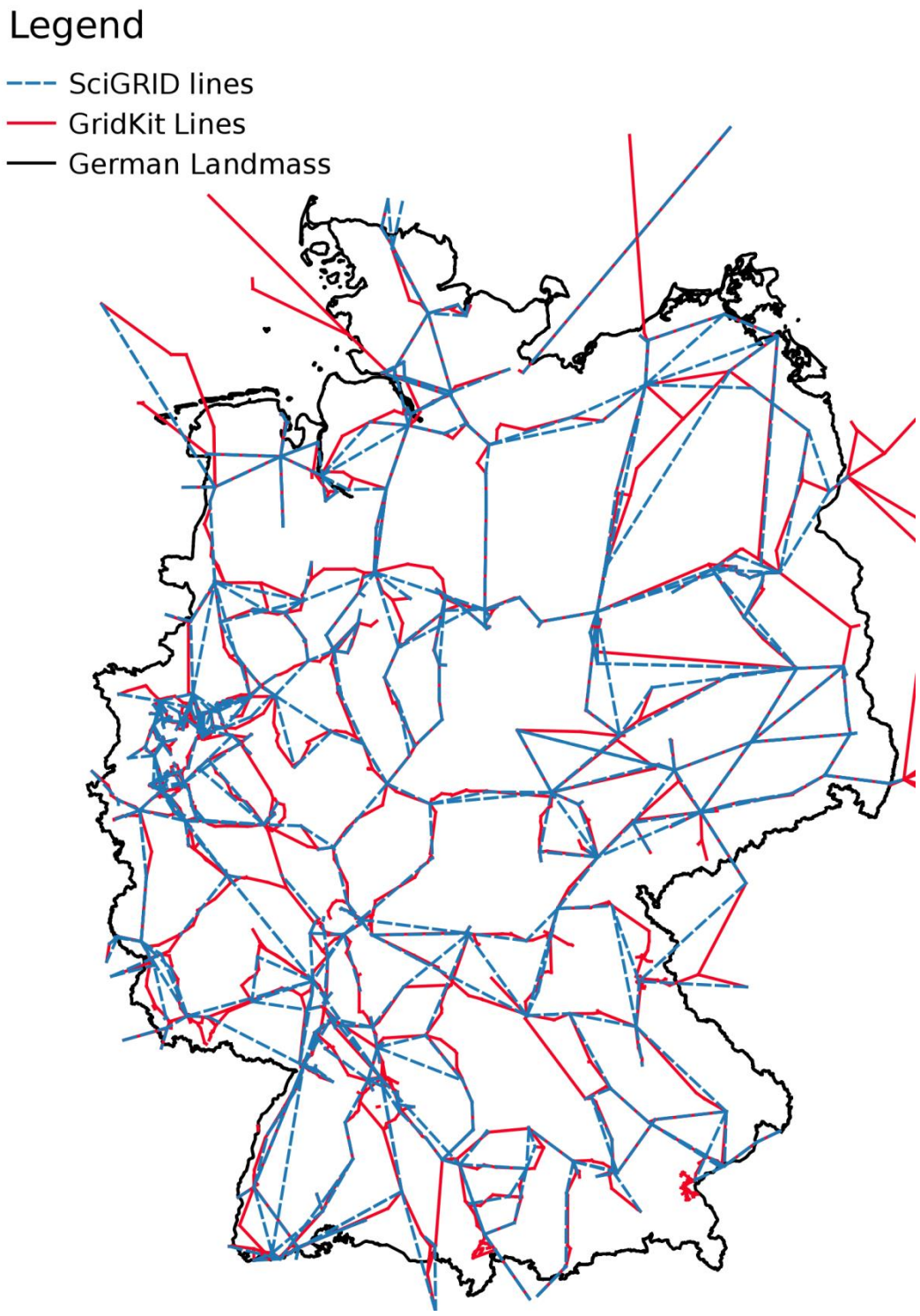

Figure 10 – German high-voltage network abstracted from OpenStreetMap, both by GridKit (red) and SciGRID (blue). The GridKit lines extend into Poland, which has been cut off in this map.

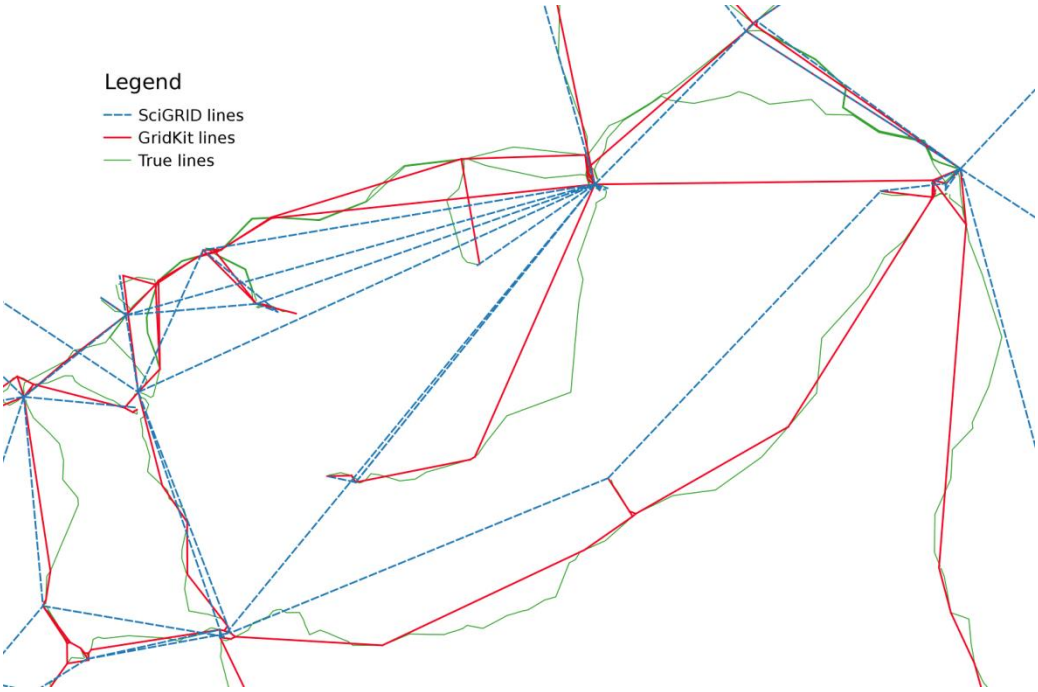

Figure 11 - Complexity of SciGRID and GridKit networks compared with the true extent of the power lines. Note that the GridKit lines follow the original 'serial' topology, whereas the SciGRID lines appear to radiate out of a central station.

The SciGRID network is much more simplified than the GridKit network. This is demonstrated clearly in Figure 11, which compares SciGRID and GridKit with the original power lines in the OpenStreetMap database. As can be seen, the GridKit network follows the original mapped lines much more closely. The SciGRID network implies a 'radial' configuration around the central station, but this is not in fact how these lines are mapped. The true configuration is much more like a 'serial' network than displayed in the SciGRID model.

Specific electrical information in the GridKit high-voltage network is less complete than it is for SciGRID. This is especially pronounced in the full network, where only 38% of lines have frequency information. Frequency information is significant because in Germany the railway power system runs at 16.7hz, and does not participate into the public electrical power system. Voltage information is relatively complete, however, at 92% for the high-voltage and 77% for the full network. Electrically relevant properties like resistance, reactance and thermal line limits can be estimated from reference values when the voltage and the number of conductor bundles and subconductors is known. The voltage and frequency information may in general be propagated from neighboring stations or lines, but this is much more complex for conductors and subconductors, since their amounts may change on a joint. Only 67% of high-voltage lines have both 'cables' (conductor bundles) and 'wires' information, compared to 78% for SciGRID. Full statistics are presented in table 3. Note that the full length of all power lines is 126.136 km, indicating that only 66% of all lines are part of the well-formed network.

From the completeness information it is possible to compute the 'relative completeness' of information for each property. This indicates the chance that a line or station will have complete electrical information given that a certain property is set. For stations, the only required properties are voltage and frequency. This indicates that for stations, the frequency value is usually the last to be added. For lines, the frequency information is often missing as well. However, it is not necessary to lookup reference values on resistance, reactance and conductance, whereas voltage level and conductor structure is. Also, frequency information can frequently be taken over from neighboring stations and lines. Thus, it was not taken into account to compute the 'completeness' of lines. With that said, wires (subconductors) has the highest relative completeness percentage, indicating that it

is usually the last to be added. This makes sense given that subconductors are not a visible aspect of power lines, and typically unmarked as well.

Table 4 - Chance that an object is 'complete' given the electrical features it contains.

Relative Completness	SciGRID	GridKit HV	GridKit (full)	Average
Stations given voltage	74%	60%	55%	63%
Stations given frequency	100%	98%	92%	97%
Lines given voltage	79%	72%	65%	72%
Lines given cables	81%	72%	66%	73%
Lines given wires	99%	97%	96%	97%

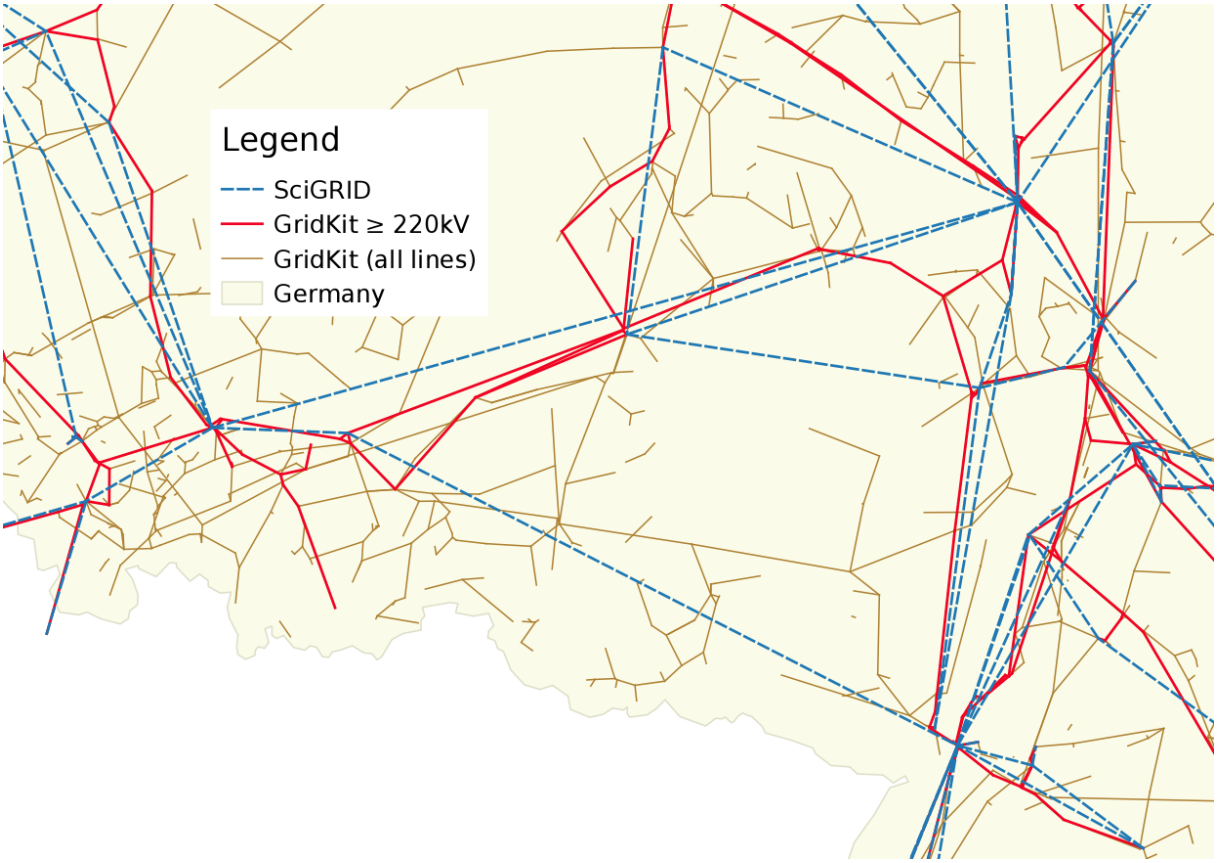

Figure 12 - A line from SciGRID that appears to be missing from GridKit (bottom left to right). While there is a network connection between the to stations linked by this line, it is not marked as a high-voltage line.

In some cases, lines are present in SciGRID, and seemingly equivalent lines are also present in the full network top ology. However, these lines are not found in the high-voltage network. See for example Figure 12. This probably indicates that the voltage and frequency information for each line was not sufficient to include it in the high-voltage network. The fact that SciGRID has found the line probably indicates that the power relation object which aggregates these lines must have this information instead.

3.2 Path Equivalence

As a result of the increased complexity of the GridKit network compared to the SciGRID network, direct neighbors in SciGRID are typically not direct neighbors in the GridKit high-voltage network. The number of 'hops' (stations) in the GridKit network between the equivalent stations of direct

neighbors in SciGRID measures this added complexity. The median number of hops is 4 and the average number is 4.89. This number is not unexpected as the GridKit model also contains 3.7 times as many lines as the SciGRID model. In some cases, there may not be an equivalent to the direct path available, which probably explains the upper end of the distribution of hops.

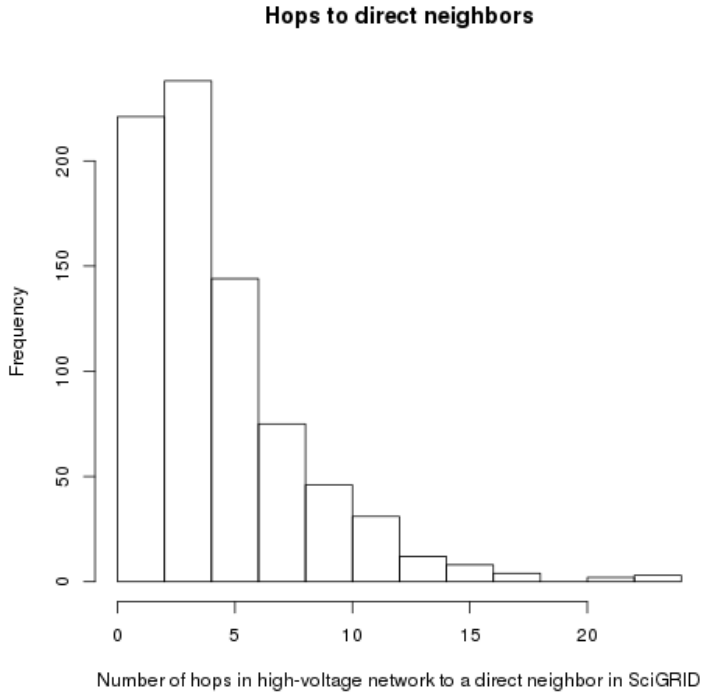

Figure 13 - Number of lines in GridKit between direct neighbors in SciGRID. The upper end of this distribution presumably represents lines missing from GridKit.

73% of paths between direct neighbors are within 5% of equal lengths in both the SciGRID and the GridKit network. There are however many outliers, especially in the direction of longer lines in SciGRID. As a result, the mean relative length is 1.07, meaning that on average the SciGRID lines are 7% longer. (The median relative length is 1.014, though, indicating that this mean is heavily skewed by the longer lines). The maximum relative length is 17.7; the minimum is 0.012. This minimum is almost certainly a result of an error in the calculation of lengths in SciGRID, since the stations involved are nearly 10 km apart while the length of the path is 662 meter.

The large number of longer paths between direct neighbors in SciGRID probably results from two factors. First of all, there are redundant lines in SciGRID that describe different paths between the same two stations. These paths are of different length, but the GridKit network always selects the shortest path. Thus the longer paths are always compared with the shortest paths. The second factor is that GridKit cuts off the section of power lines that lie inside a power station. This makes them appear shorter than they really are. On average, a station has an area of 111.000 m^2, equivalent to a square of 330 meter on each side. A line that might extend from the center of one such station to the center of another would appear to be up to 330 meter shorter. This may represent a significant fraction of its length, particularly when the lines are close together.

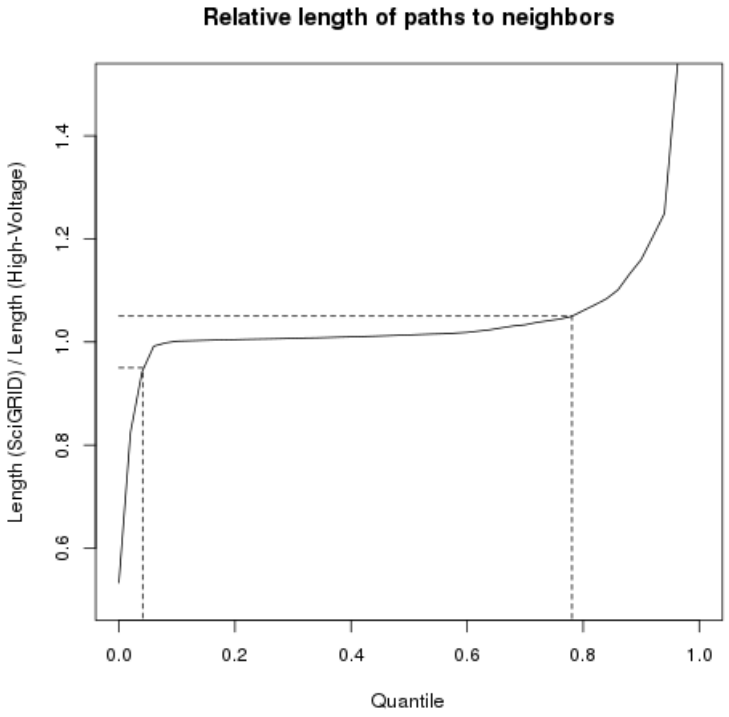

Figure 14 - Relative length of paths in GridKit to direct neighbors in SciGRID. Most values are close to one, which implies that the paths through both networks are equivalent. Lower values may indicate a missing path in GridKit. Values greater than one can be attributed to line length distortion as well as possible redundant paths in SciGRID.

The fact that most relative path lengths are close to 1 strongly suggests that both models use the same lines and that both methods of line length calculation are compatible. If they were not (for example due to different projection systems) a systematic deviation from 1 would be expected. Therefore this rules out both different methods of line length calculation or a radically different network structure as explanations for the considerable difference in total network size between SciGRID and the GridKit high-voltage network.

In total 110.350 paths from every cross-identified station to every other cross-identified station have been computed for both network. In contrast with direct neighbors, only 25% of lines are within 5% of equal length. The mean relative length is 1.25, meaning that on average a path through the SciGRID network is 25% longer than in the GridKit high-voltage network. This plausibly indicates that there are lines which are present in the high-voltage network which are absent in the SciGRID network, and that these missing lines contribute to the overall structure of the network. An extreme example is the case of two stations in Audorf, Germany that lie 120 meters away from each other. Because the line connecting them is missing from SciGRID, the shortest path between them is 130 km long.

The relative number of hops between equivalent paths is somewhat less in the general case (3.28 mean relative hops, 3.16 median) than it was in the case of direct neighbors. It is even somewhat less than the ratio between lines. Together with the relative length of equivalent paths – which are on average longer in SciGRID than in the high-voltage network - this strongly suggests that the high-voltage network is able to use paths which are unavailable to the SciGRID network. Thus, it suggest that the high-voltage network is not only more complex but also topologically more complete than the SciGRID network.

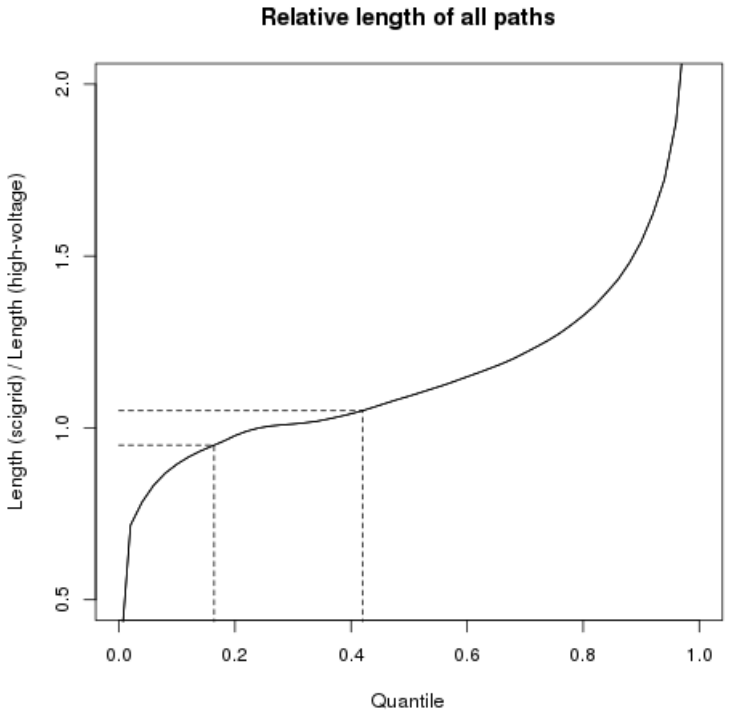

Figure 15 - Relative length of paths between any two stations in that are present in both SciGRID and GridKit. Nearly 60% of these paths are more than 5% longer in GridKit than in SciGRID whereas around 20% of paths are shorter in SciGRID.

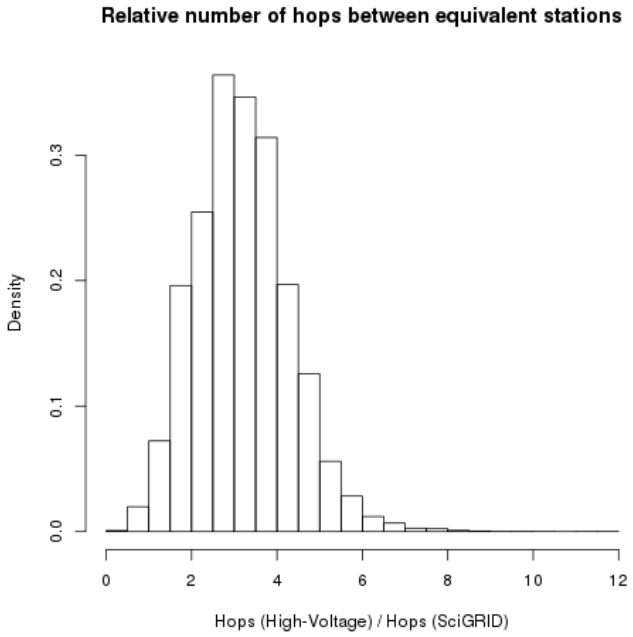

Figure 16 - Relative number of hops between equivalent paths in SciGRID and GridKit networks

3.3 Application beyond Germany

In order to determine if the heuristic method used in GridKit is applicable beyond Germany, the abstraction method has been applied to an OpenStreetMap extract from the European area. Figure 17 shows an extract drawn on top of the official 2014 ENTSO-E map of the European power grid, limited to France. (Showing the whole map would leave the ENTSO-E lines invisible). It demonstrates quite good agreement between the official map and the network extracted from OpenStreetMap. However, no comparable network is available for this wider area, so it is impossible to validate this further. Table 5 lists a number of power statistics for power lines in the European area.

Figure 17 - Northern France and Southern Belgium, drawn on top of the 2014 ENTSO-E grid map of Europe. The ENTSO-E map is not expected to be perfectly accurate. Agreement in this area is good, although not perfect.

Table 5 – Statistics of GridKit extracts of various European countries.

Country	Number of lines	Total line length	Average line length	Lines with voltage
Belgium	140	1.467 km	10,5 km	98%
Luxembourg	38	169 km	4,4 km	100%
Austria	239	3.929 km	16,4 km	85%
Denmark	130	2.142 km	16,5 km	99%
Switzerland	639	5.573 km	8,7 km	65%
Netherlands	130	1.731 km	13,3 km	100%
Germany	2853	20.487 km	7,2 km	91%
France	2399	37.139 km	15,5 km	90%
Europe	19352	317.144 km	16,4 km	67%

Further examples are provided by Figure 18 and Figure 19 for the Netherlands and central and east of the USA.

Because the Netherlands is a small country, few high-voltage lines are present, and most areas are supplied by medium-voltage (110kV) lines. In contrast, in the USA the only lines shown by the official EIA map are 345kV and higher.

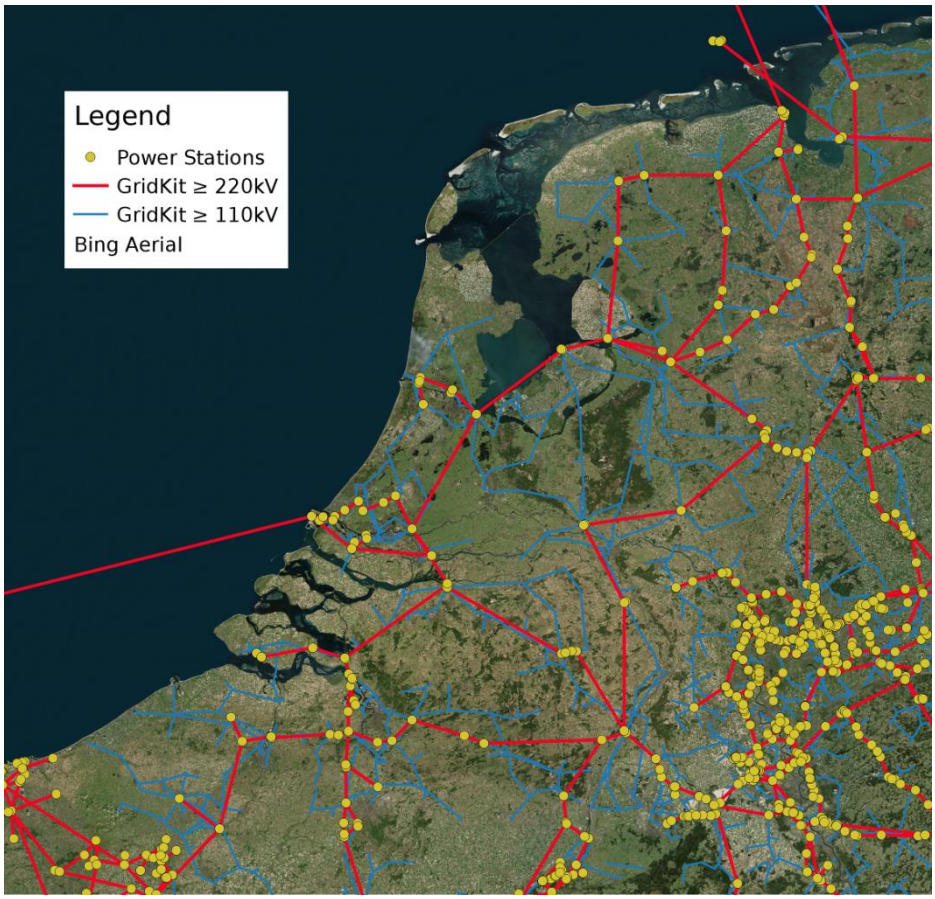

Figure 18 - Dutch high and medium (110kV) voltage power grid, extracted by GridKit.

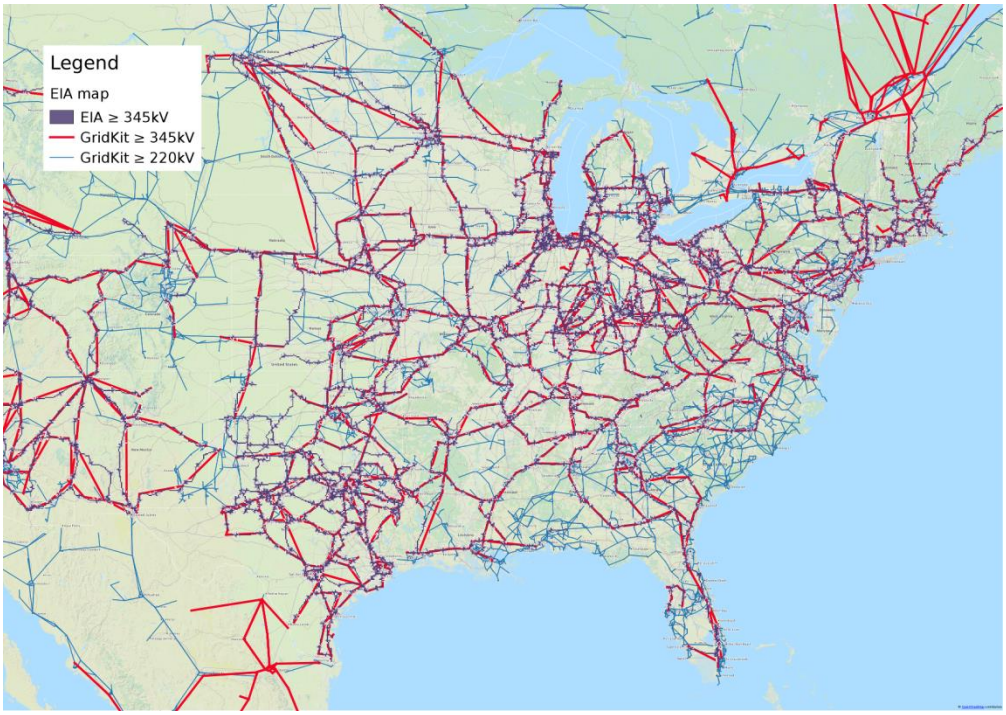

Figure 19 – East coast and central USA, overlaid on a grid map provided by the EIA

4. UNCERTAINTY, VALIDITY, SENSITIVITY

The fact that the total length of all lines in SciGRID is longer than in the GridKit German high-voltage network presents a mystery. The reason why the SciGRID database extract contains the Polish network as well as the German network is easily explained, because it is an artifact of the extraction process from the full OpenStreetMap database. SciGRID relies on 'relation' objects in the source database, so during the filtering process it retains all power relation objects which overlap at least partially. The full extent of Polish high-voltage lines have been collected within a single relation and contain some lines which cross the border with Germany. Thus, all high-voltage power lines in Poland are retained in the data extract used by SciGRID. However, specific Polish power routes have not been retained in this filtering process, and thus the SciGRID model does not contain them.

As mentioned above, differences in line length calculation (for example due to the use of incompatible projection systems) are unlikely because direct neighbors have typically equivalent paths with nearly equal lengths. This would be very unlikely if all line lengths were distorted in some way. However, a clue can be taken from the radial topology shown in Figure 11. Figure 20 shows another example of this phenomenon. In each of these cases SciGRID derives a direct topological relation between stations which are arranged in a series. This must be because there is a 'relation' object which describes the connection between each of these stations. This relation also describes the path which is used between each station. It is the length of this path which is reported as the line length in the SciGRID network.

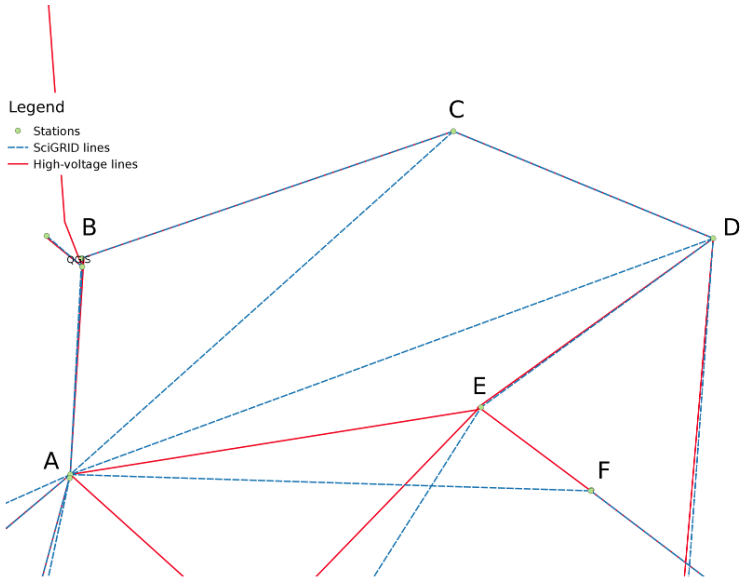

Figure 20 -A radial topology in a serial network. The true path from station A to C passes over station B, and it is plausible that the true path from station A to D passes over station E. Hence the lines A-B, B-C, A-E and E-D are all counted more than once.

However, because these stations are really arranged in a series, this means that the paths between the station which are counted as distinct lines in SciGRID are *shared* by multiple paths in reality. Thus, the paths between A and B and B and C are counted more than once, which causes the total length of lines to be overestimated. Because SciGRID does not store many intermediate results, it is difficult to determine exactly the extent of this overestimation. But a preliminary analysis finds up to 1344 lines which are shared in multiple relations. The total excess line length could be as much as 15.000 km, which leaves the SciGRID network at 15.500 km in total. This would be consistent with the higher degree of connectivity in the GridKit high-voltage network. However, it is very difficult to determine the extent to which these duplicate lines are really used to form the SciGRID network. Also, SciGRID

considers every set of 3 cables to be a single path for the consideration of line length. If a certain line has 6 cables which are shared by two relations, then each of these two relations may acquire 3 cables, in which case the line length double-counting is intended. No such 3-cable logic is applied to the counting of line length in the GridKit model, which may also partially explain the difference.

The derivation procedures used by GridKit are highly sensitive to buffer sizes. Buffers enable the discovery of nearby objects by extending the area around the object and searching for things that lie within that buffer area. Using buffers that are too large causes false positive matches, whereby stations and lines intersect accidentally. This is especially a problem for smaller 'substation' objects, since this power type is applied to transformer boxes and industrial installations equally. There is a 'substation' tag defined which should allow for further specification of station type, but it is rarely used and cannot be relied on. However, using smaller buffers will cause many connections not to be found. Thus, a balance has to be struck. Initially the buffers used were 150 meter around stations and 100 meters around line terminals. These values were far too large and in the final version 100 and 50 meters are used as maximum values. The optimal size of buffers will probably depend on the specific characteristics of the dataset that is being processed.

No method has been applied for the elimination of false connections. Originally it had been envisioned that it would be possible to find impossible electrical combinations that had been introduced by the abstraction procedures, and remove them from the network. But seemingly impossible combinations – like power lines that carry different frequencies – are mapped out in the original data, and are presumably correct. This leaves an automated procedure with rather little authority to eliminate lines or stations because of such combinations. This means that the burden of correct interpretation of electrical properties on such lines is then on the users of the network model.

SciGRID may not be an ideal reference model. Ultimately it is based on the same OpenStreetMap database as the GridKit network. The relation objects that it contains are human-defined, and this provides them with some authority. However, it may also mean that the same errors in the source database are compared to each other. Using SciGRID as a reference model cannot prove that the GridKit network properly models the real world, because this is unproven for SciGRID. Furthermore, SciGRID is (currently) limited to the German network. If the GridKit high-voltage network matches closely to the SciGRID network this does not imply that it will match closely to the network of other countries, even though this would seem to be visually confirmed.

After this research was started a geo-referenced version of the 'Bialek' model, which is based on the power system maps that are made public by ENTSO-E, has become available (Jensen et al, 2015). Comparing this model to the GridKit network would provide a better validation. However, it is challenging to compare these two models, since there is hardly any spatial overlap between nodes. In Germany, the distance between nodes that appear to be equivalent (based on their connected lines) can be over 40 km. Thus, another approach is necessary to determine node equivalence between both models. Because the GridKit high-voltage model for the same area (Europe) has an order of magnitude more nodes (14.349 vs 1.494) such identification might not be possible to automate.

The method for comparing topologies based on shortest paths that was newly developed and used for this research. General methods for comparing similar but unequal networks could not be found. While this method is intuitively appealing, it has not been formally or empirically validated. A specific bias of this method is that it only accounts for missing lines – it rewards falsely added connections by making average paths shorter. Whether or not that is a common occurrence is questionable. This comparison method also does not take into account the electrical properties of the lines. In comparing electrical networks the least resistance path is probably more interesting than the shortest spatial path. Computing the least-resistance path is no more complex than computing the spatially shortest path.

5. DISCUSSION

This work has demonstrated the feasibility of extracting a well-formed high-voltage network from volunteered geographical information recorded in the OpenStreetMap database. In contrast with earlier work, this method relies only on spatial features and electrical information recorded on an object level. Such features are available by definition – otherwise they would not be drawn on the map as power structures and mapping them would be pointless. Because of this, it is practically universally applicable wherever elements of the power system have been mapped. It has also been demonstrated for the power networks of Europe, the USA, and elsewhere (not shown here).

The core assumption underlying the design of the procedures is that maps are drawn rather than laid-out. Thus, a significant amount of work has been expended to transform power lines and stations from their complex, intricately drawn shapes and forms into simple direct lines and polygons that a computer is able to process. For a very large proportion of the network – well over 98% of lines – these procedures are effective. However, there remains a large proportion of all mapped lines – up to 34% in Germany, and more in other countries – which cannot be connected to at least two stations, meaning they play no part in the topology. In other words, a large part of the electricity network is mapped only as individual elements that are not connected to any other part.

A project initiated in 2015 at the Fachhochschule Flensburg also attempts to construct a model of the high-voltage transmission system in Germany, named *osmTGmod* (Scharf, 2015). The method used by osmTGmod is hybrid between the one applied by SciGRID and the one proposed here. First the 'power routes' are extracted from the OpenStreetMap database. Then, after removing these lines and all lines which cannot be validated in some way from the original set, a general purpose routing algorithm is applied to the remaining, sparse, network. (Scharf, 2015). This hybrid approach may benefit from both the presence of (authoritative) power routes as well as from the more primitive individual items. On the other hand, the method applied by *osmTGmod* is pessimistic in nature, because it relies on removing as many items from consideration as possible. It also uses a very large buffer size (300 meter) with the routing algorithm. In a very sparse network a large buffer size will likely not be as problematic as it is in a dense network. However, may not be possible to apply a prior pruning step in networks outside Germany.

During the extraction of this model a large amount of intermediate results are generated. These intermediate results provide insight into how power system structures are processed and enable the validation of results. They can also be used to provide feedback to the OpenStreetMap community and point them to specific map objects which might be improved. For instance, it may be possible to extend the 'dangling' lines and isolated stations to connect to the greater network, or to identify direct lines which are possibly transmission lines but which are not sufficiently well 'tagged' to allow this inference. In short, intermediate results and rejected items provide actionable feedback for map improvements.

A challenge to any method of automated extraction of the power grid is the limited resolution of power structure types. The 'substation' type is applied both to large installations and small street transformer boxes. (There is a power type 'transformer', but it is rarely used). It is advisable to introduce new and more specific power types that allow such structures to be distinguished. This can be as simple as proposing the change and adding it to the OpenStreetMap wiki. However it is unknown how long it would take for changes in the tagging of these structures to propagate and whether this would ever provide sufficient reliability.

In this work and in SciGRID it is assumed that the electrical properties (resistance, reactance, conductance and thermal current limits) of power lines can be computed from reference values once the voltage and structure of conductors are known. This assumption relies on the idea that all power lines are built in the same fashion from the same materials. While this may be reasonably accurate in one country (especially one with a modern infrastructure) it is not necessarily accurate for all. This is

especially challenging when applying the network abstraction procedures to a larger network, such as continental Europe. It might be advisable to implement region-specific reference values in such cases. This type of electrical information has so far not been recorded in the OpenStreetMap database and it is unlikely that it will be, making the use of reference values a necessity.

The network model that is extracted by this work can be used in power flow simulations. However, it contains topological information only. It does not contain generation and load information which would be used in a practical application of a simulation. Without this information, power flows will often be impossible to compute. Thus it is likely that this will prove to be a stumbling block for researchers who would lke to investigate the effect of specific changes, e.g. the introduction of a new HVDC line or a wind turbine installation. This should be addressed by integrating information from third-party sources. For example the Enipedia website collects power generation information of a large number of power plants from a multitude of online databases. Enipedia acts as an aggregate database that can be automatically queried. Hence it may serve as a source for reference generation information that can be published together with the network topology.

Compared to other network models such as that provided by Bialek (Bialek and Hutcheon, 2009) or SciGRID (Medjroubi and Matke, 2015a) the GridKit topology of the same area is considerably more complex, with a ratio of 4 lines to 1 in Germany and 10 to 1 in Europe. It is probably also considerably more accurate, at least geographically. This follows from the use of actual line extents and precise station locations, and the consistent introduction of joints wherever lines split or combine. But this geographical and topological accuracy may be of limited value to researchers, while the additional complexity can be costly. Calculating power flows over 1.500 nodes may be feasible when calculating flows over 15.000 might not be. This is especially true when these calculations have to be performed repeatedly, for example in a time-series simulation. It is likely that this complexity can be reduced considerably in many cases with a moderate amount of distortion. This is essentially a graph-optimization problem, for which many techniques are known. However, such techniques can often be complicated to implement.

This work has also demonstrated a method of comparing distinct networks in which a reasonable subset of equivalent nodes can be identified, by means of comparing shortest-path lengths. It is unlikely that such a straightforward method is completely novel. However, it has been very challenging to find prior work describing this strategy. The requirement that equivalent nodes can be identified may in some cases be hard to meet, even with spatial information. A full network flow calculation simulation would have provided a more convincing validation of the extracted network. Results of such simulations could even be compared to historical data, as is done by Bialek and Zhou (2005). However, as mentioned above, this requires acquiring accurate per-station load and generation data, which is challenging in its own right. Furthermore from the results of power flow simulations topological errors cannot be distinguished from errors in power load and generation information. The method of comparing shortest paths does allow for finding purely topological errors. The statistics of these calculations describe the system, but it can also be meaningful to investigate the individual paths for large anomalies.

6. CONCLUSION

A well-formed network of high-voltage power lines, power plants and substations can be effectively be extracted from open data in the OpenStreetMap database, using only spatial and electrical features on individual elements. The resulting network model is considerably more complex than comparable models, but it is also complete and applicable to almost any area where power system structures have been added to OpenStreetMap. This includes most parts of western Europe. As a side effect, it can be used to supply feedback to OpenStreetMap itself by indicating problems in specific power lines and stations.

This work thereby enables the study of the role of the electric power grid in the integration of and transition to renewable energy. A significant but not insurmountable challenge remains in acquiring or estimating load and generation information for use in power flow calculations. Only with accurate information on the power supply and demand can power flow calculations be performed, and with that the network topology can be validated. This is true both of SciGRID as well as the GridKit network model.

The results demonstrate that while volunteered open data may not always be consistent, it can be surprisingly complete and useful. By means of relatively straightforward procedures a large variety of structures can be simplified into simple elements that are suitable for automated analysis. With an optimistic, careful and case-specific approach open data can be a highly powerful instrument in energy research.

7. REFERENCES

Zhou, Q., & Bialek, J. W. (2005). Approximate model of European interconnected system as a benchmark system to study effects of cross-border trades. *Power Systems, IEEE Transactions on*, *20*(2), 782-788.

Hutcheon, N., & Bialek, J. W. (2013). Updated and validated power flow model of the main continental european transmission network. In *PowerTech (POWERTECH), 2013 IEEE Grenoble* (pp. 1-5). IEEE.

Matke, C. and Medjroubi, W. (2015-a) SciGRID user guide v0.1. Published at http://scigrid.de/releases_archive/SciGRID_Userguide_V0.1.pdf.

Matke, C. and Medjroubi, W. (2015-b) Power relations in SciGRID. http://scigrid.de/posts/2015-Jul-02_power-relations-in-openstreetmap.html. Accessed at 2015-11-09.

Semerow, A., Hohn, S., Luther, M., Sattinger, W., Abildgaard, H., Garcia, A. D., & Giannuzzi, G. (2015). Dynamic Study Model for the interconnected power system of Continental Europe in different simulation tools. In PowerTech, 2015 IEEE Eindhoven (pp. 1-6). IEEE.

OpenStreetMap Wiki: Node. http://wiki.openstreetmap.org/wiki/Node Accessed on 2016-01-12.

Wikipedia: World Geodetic System. https://en.wikipedia.org/wiki/World_Geodetic_System. Accessed on 2016-01-12.

Wikipedia: Map Projection. https://en.wikipedia.org/wiki/Map_projection. Accessed on 2015-12-18.

ArcGIS Blog: Measuring distances when your map uses the mercator projection.

https://blogs.esri.com/esri/arcgis/2010/03/05/measuring-distances-and-areas-when-your-map-uses-the-mercator-projection/. Accessed on 2015-01-05

Jensen, T. V., de Sevin H., Greiner M. & Pinson P. (2015). The RE-Europe dataset.

https://zenodo.org/record/35177#.VsEiz0KCCV4. Accessed on 2016-02-14

Spatialite Wiki: Graphs Intro. https://www.gaia-gis.it/fossil/spatialite-tools/wiki?name=graphs-intro. Accessed on 2016-02-15.

Wikipedia: Kirchhoff's Circuit laws. https://en.wikipedia.org/wiki/Kirchhoff's_circuit_laws. Accessed on 2016-02-15.

Enipedia Power Plants: http://enipedia.tudelft.nl/wiki/Portal:Power_Plants. Accessed on 2015-12-09.

Principles of Sustainability. Chapter 6: Energy Sustainability; http://www.webpages.uidaho.edu/sustainability/chapters/ch06/ch06-p3a.asp. Accessed on 2015-03-02.