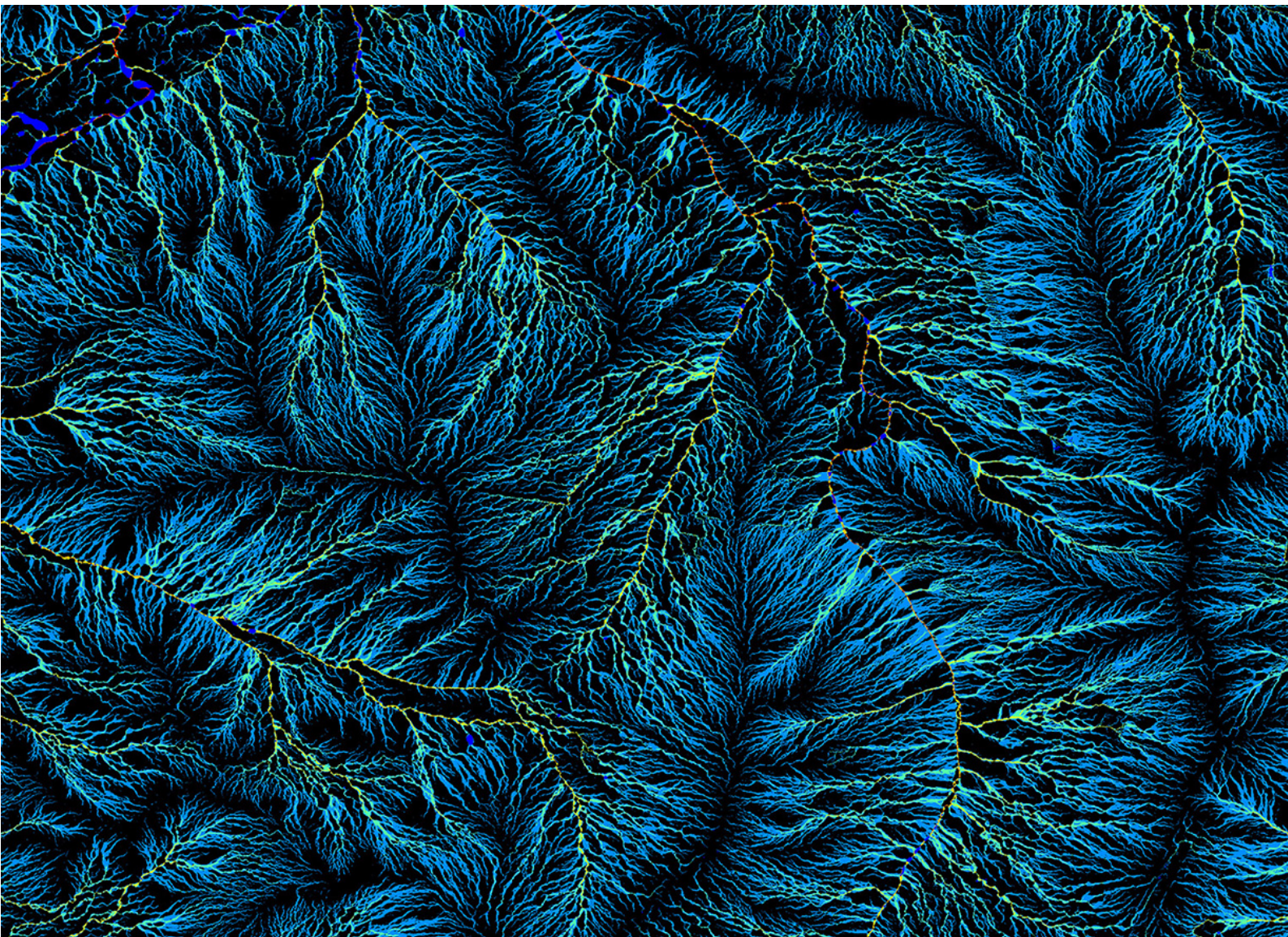
Australia's National
Science Agency

National Low-Voltage Feeder Taxonomy Study

Public Release Version 1.1.2

October 2021



Citation

Geth F, Brinsmead TS, West S, Goldthorpe P, Spak B, Cross G and Braslavsky J (2021) National Low-Voltage Feeder Taxonomy Study. CSIRO, Australia.

Copyright

Important disclaimer

Contents

Acknowledgements

This project received funding from ARENA as part of ARENA's Advancing Renewables Program. The views expressed herein are not necessarily the views of the Australian Government, and the Australian Government does not accept responsibility for any information or advice contained herein.

Acknowledgement of funding

The National Low-Voltage Feeder Taxonomy Study is led by the Commonwealth Scientific and Industrial Research Organisation is an Australian Government agency responsible for scientific research. The CSIRO works with leading organisations around the world.

The National Low-Voltage Feeder Taxonomy Study received funding from ARENA as part of ARENA's Advancing Renewables Program.

Acknowledgement of project partners

The CSIRO would like to acknowledge the following project partners: Energy Networks Australia, Ausgrid, AusNet Electricity Services, Western Power, Endeavour Energy, Energy Queensland, Essential Energy, Horizon Power, SA Power Networks & TasNetworks.

Acknowledgement of technical assistance

The authors would like to acknowledge the technical assistance and support of Matt Amos with software integration, Giles Browne with geospatial data acquisition and analysis, Rahmat Heidarihaei with power flow analysis, Terijo Lovasz with editing, and Anson (Ching Hong) Tam with load data search and analysis.

Project summary

In order to improve the hosting capacity of distributed energy resources (DER) across Australia, a greater understanding is required of low-voltage networks, to allow design and assessment of appropriate technologies and systems.

The National Low-Voltage Feeder Taxonomy Study is the first national low-voltage network taxonomy that outlines the real-world characteristics of Australian low-voltage distribution systems.

In doing so, we have developed a realistic, publicly available, dataset and models describing the most common types of low-voltage networks found in Australia.

The purpose is to inform and facilitate early exploration in simulation of new energy technologies – it does not avoid the need for assessment for deployment in specific scenarios experienced in different distribution network areas. These models will enable users to test the value proposition of innovative technological solutions by highlighting how they contribute to the stability, reliability, and performance, of networks across Australia.

The project has also identified specific areas where we need to improve the understanding of the challenges and opportunities associated with using data to improve DER integration and encourage greater overall utilisation of the network.

The 23 identified representative network models plus associated workbooks can be downloaded now as: **'Low-Voltage Feeder Taxonomy Study - Data & models'** *release hosted on the Australia's National Energy Analytics Research Program website (NEAR) website. Download via* https://near.csiro.au/assets/f325fb3c-2dcd-410c-97a8-e55dc68b8064 *.*

Background

In early 2019, ARENA announced it was providing $9.6 million in funding to support 12 projects focused on integrating small-scale solar, batteries, electric vehicles and other responsive consumer devices (also known as distributed energy resources – DER). As ARENA CEO, Darren Miller, said in the accompanying press release, the projects focused on 'how we can make the most of the growth in distributed energy resources'[1].

Many of the projects focused on addressing a primary challenge that continues to curtail the full utilisation of DER: identifying and communicating the status of the network so that DER can operate in a way that respects – and even helps address – network limits. Indeed, ARENA funded CSIRO to 'identify a concise set (taxonomy) of low-voltage (LV) network types and associated

[1] https://arena.gov.au/news/distributed-energy-projects-awarded-nearly-10-million/

models, with the aim of facilitating consistent and effective DER hosting capacity analysis by DNSPs, researchers and other stakeholders.'

The importance of effectively identifying network limits – often referred to as the network's 'hosting capacity' – has become even more central to the discussion of DER integration in the months since the projects were announced. In March 2021, the Australian Energy Market Commission (AEMC) submitted a ground-breaking draft determination 'to clarify that distribution services are two-way and include export services.' The need to change the regulations for DNSPs was, according to the AEMC, prompted because networks are 'approaching the limit of their "intrinsic hosting capacity".'[2]

In addition to the AEMC, others have identified the need for the greater transparency and consistency in identifying hosting capacity. For example, the Energy Security Board's Data Strategy consultation paper recommends 'Requiring networks [DNSPs] to publish their estimated DER hosting capacity to help inform investments and decisions around DER connection requirements'[3]. In late 2019, an AER-commissioned study authored by the CSIRO and Cutler Merz to identify methodologies for determining the valuing of DER (VaDER) recommended that the AER develop 'guidance for networks [DNSPs] to follow in assessing the hosting capacity of their networks.'[4]

The Project's Aims and Achievements

The Low-Voltage Feeder Taxonomy (LVFT) project had two primary aims. The first was to develop a realistic, publicly available dataset and models describing the most common types of low-voltage networks found in Australia. The second was to improve understanding of the challenges and opportunities associated with using data to improve DER integration and encourage greater overall utilisation of the network.

This report focuses primarily on the first objective; CSIRO partnered with several DNSPs to collect and analyse existing network data from more than 90 thousand low-voltage networks to identify an archetypal set of 23 low-voltage networks and build associated models of these networks demonstrating the impact of DER operation on various network metrics.

To a certain extent the project's more ambitious objective, to develop a national taxonomy representative of **real-world distribution networks**, was a failure – instead we developed a taxonomy of **distribution network data**, based on the incomplete data that the DNSPs we collaborated with have assembled it so-far. The primary challenge in developing network models that are truly representative of network conditions is that DNSP businesses lack reliable impedance and phase connectivity data that would allow the models to accurately represent real-life conditions. As we do not know how representative the power flows and congestions are for the real world, this taxonomy should be used as a testbed only, not as a basis for extrapolating the hosting capacity of Australian distribution networks.

---

[2] https://www.aemc.gov.au/sites/default/files/2021-03/Draft%20Determination%20-%20ERC0311%20and%20RRC0039%20-%20Access%20Pricing%20and%20Incentive%20arrangements%20for%20DER.pdf

[3] https://energyministers.gov.au/sites/prod.energycouncil/files/publications/documents/ESB%20Data%20Strategy%20Media%20Release.pdf

[4] https://www.aer.gov.au/system/files/CSIRO%20and%20Cutler%20Merz%20%E2%80%93%20Value%20of%20distributed%20energy%20resources%20-%20Methodology%20study%20%E2%80%93%20Final%20report%20%E2%80%93%20October%202020_1.pdf

With that said, this set of networks and their associated models should provide useful tools to researchers and technology companies that require baseline data for testing different approaches for improving DER integration. The purpose of a testbed is to inform and facilitate early exploration in simulation related to the integration options of new energy technologies. In the absence of more accurate data on Australia's low-voltage networks, the enclosed set of network models is likely the best publicly available dataset.

The second primary project objective – better understanding the challenges and opportunities associated with network data – has proven to be a fruitful area for initial research. In short, there is a sizeable gap between the amount of data, tools, skills and capabilities required to enable a low-cost bidirectional grid and the current amount of data and digital capability that exists in the sector. The key learnings and opportunities for improvement can be categorised into four separate topics: i) the role of network models; ii) improvements in data collection and cleaning; iii) digital skills and capability; and iv) data sharing. Before touching directly on these areas for improvement, however, it is important to underline the importance of data to the energy transformation underway in Australia.

The importance of data in the electricity transformation

Industry commentators regularly discuss the various 'D's of the energy transformation: decarbonisation, decentralisation, democratisation and digitalisation. There is an important distinction between these underlying trends currently dominating the industry, however. While decarbonisation and decentralisation/democratisation are either policy goals or consumer-directed outcomes, digitalisation is a tool. Digitalisation is not a required element of an electricity system – indeed, the dynamo preceded the computer by several decades. Decarbonised and decentralised electricity systems do not require greater use of data and digital technologies; microgrids and hydropower were likewise early favourites at the dawn of the electricity industry.

Nevertheless, an *affordable and modern* decarbonised and decentralised electricity system surely requires greater deployment of digital technologies. The natural monopoly of electricity infrastructure frustrates, if not disallows, the revolution that digital technology firms have affected in other sectors, such as Uber in transport, Airbnb in hospitality, voice over internet (VoIP) in telecommunications, and Amazon in retail. And yet consumers expect similar revolutions in the electric service they receive.

Traditionally, a unidirectional electricity grid – with power coming from a few, large, faraway power plants – delivered affordable and reliable energy through a 'set and forget' design. Such a grid can be well designed largely through effective planning, and relatively static customer behaviour enabled planning based on reliable assumptions. Technology, specifically low-cost solar panels, changed all that.

Roughly one in four Australian households have solar today, and all signs indicate that customer adoption of solar is still accelerating. Add batteries and electric vehicles – whose widespread adoption has not yet occurred but can be reasonably assumed to be imminent – and there results a very dynamic grid, with significant amounts of power coming from what was previously the 'end of the line,' and an increasingly large range of behaviour from customers. In short, the grid now changes so much, so quickly, that relying almost exclusively on planning and a 'set and forget' design is no longer fit for purpose.

The logical alternative is to improve and increasingly rely on operations – that is, actively managing the grid in short timeframes (making regular changes to settings at least weekly or daily, if not every hour or every minute). Dynamic operations and management rely on data and reasonable visibility of the grid's status.

As the Energy Security Board Chair, Kerry Schott, has remarked, 'We now have more data than ever, but it isn't being fully utilised and shared in a way that benefits consumers or provides the information necessary to inform investment in Australia's energy future.'[5]

The role of models in DER Integration

There is an ongoing tension within the Australian electricity industry to downplay the value of network models themselves. This resistance to network models comes in a variety of forms. Many influential but uninformed commentators often remark that the data challenge in integrating DER is exclusively related to difficulty in deploying and accessing smart meter data and data from DER devices and inverters. While there is no doubt importance and value in improving end-point device and customer data, these commentators dismiss outright or significantly downplay the need and value for network topology data – i.e. data that describes the geometry of the network, showing the location of customers to one another and to transformers, line impedance, etc. Such data is critical to accurately depicting the physics that ultimately govern the operation of electricity infrastructure.

Those closely associated with DNSPs and policymakers often recognise the value of network topology data but insist that it is too challenging or expensive to collect and therefore not worth the effort. They argue that the ability of some DNSP businesses to communicate network constraints – as several have done in various ARENA trials – demonstrates that existing data collection efforts are more or less satisfactory. And indeed, there is potential to improve the operational performance of networks on the basis of real-time measurements alone, rather than also requiring detailed knowledge of network parameters. The challenge with this perspective is that it fails to recognise that the network constraints that are communicated are likely quite conservative, because with limited information the natural and proper course for DNSP businesses is to avoid exceeding the capacity of the network by underestimating its limits. This approach lowers the overall utilisation of the network, increasing costs to customers. Because DNSPs lack visibility of their overall utilisation of their assets – and how that utilisation may or may not change with greater data – it has proven challenging for DNSPs to create compelling business cases for collecting significantly greater data on the network.

The AEMC's draft determination on Access and Pricing hints at a possible reckoning with the need to more clearly identify network limits and the data used to calculate them. To promote greater transparency in DNSPs' provision of export services, the draft determination requires DNSPs to report on a range of (yet unidentified) metrics related to export service performance in their distribution annual planning reports (DAPR). More importantly, perhaps, the draft determination also links the provision of export services to DNSP financial incentives, stating that 'export service levels will be guided by performance targets that the DNSP will be incentivised to maintain and

[5] https://energyministers.gov.au/sites/prod.energycouncil/files/publications/documents/ESB%20Data%20Strategy%20Media%20Release.pdf

improve on'. Credibly establishing those performance targets will require a clear identification of network parameters to measure or model and the establishment of reasonable approaches for measuring and modelling such approaches.

The reality is that improvements in DER & customer data, and improved network models, are mutually reinforcing; better DER & customer data enables better models, and good network models can help augment and improve DER & customer data. Identifying the most cost-effective combinations of smart meter and DER device monitoring data, network topology data, and network modelling to support DER integration, however, is an unmet need. Without more information on what performance gaps can be filled with models, and which data is best directly monitored and captured, the industry is largely blind to what data is most worth collecting. As a result, we see an obvious opportunity for future research through extensive monitoring of a relatively small number of real networks (roughly a dozen) that have more data collected on them than would be cost effective to do throughout Australia. With strong datasets established, these low-voltage networks could serve as benchmarks for assessing the value of various technologies and models that attempt to estimate network parameters that are not directly measured. Indeed, the process of building these benchmarks can serve to further identify the gaps between best-practice network data collection, smart meter sensor data, and simulation, and therefore is crucial to the development of improved processes and workflows for improving relevant network data.

Data Collection and Cleaning

It is sometimes noted that the 'big data' revolution is just as much the advent of 'dirty data,' and the LVFT project ran into several issues with data cleaning. The largest challenge that the project faced was comparing common features across different datasets that were not described and maintained in standard ways across DNSPs.

Varying levels of network data completeness, missing distributed generators, loads and service lines, missing grounding data, missing feeder information, undocumented GIS coordinate systems, and inconsistent or missing switch labelling, impedance representations and transformer configurations were among the most problematic and prevalent features of the majority of network models provided. For all but two DNSPs, the provided network data were incomplete and biased samples of the actual low-voltage network assets owned by the DNSPs. From the perspective of this project, these data issues undoubtedly biased results towards the more complete datasets, and likely resulted in the identified clusters missing important network topologies. From a broader perspective, the lack of standardised data formats and availability of software tools to clean and parse network data file formats creates a significant barrier to innovation, and indeed to improving network utilisation.

Another challenge is that network asset data should be maintained to be consistent with the assets in the field. If lines are replaced, added, or re-configured, the augmented design be represented in all the relevant software systems: GIS, ADMS and power engineering simulation platforms such as PSS/Sincal or PowerFactory. At many DNSPs, it is not clear that any of these software solutions today represents a 'single source of truth'. Moreover, existing records are likely based primarily on the network as designed, rather than the network as built. Anecdotal evidence suggests business process changes and improvements in workflows, such as new reporting responsibilities for line crews and property developers, are needed to ultimately solve some of these challenges. Cultural and business change management innovation within DNSPs may prove

as, if not more, important and challenging as the development and application of better software solutions.

We see at least two ripe opportunities for improvements in data collection and cleaning. First, there is significant value in creating realistic, publicly available datasets of typical load profiles for customers. These datasets should include fast time resolution (e.g. one-minute), and at least separate active and reactive power profiles. AEMO's NMI data contains only real power data with a 30-minute time resolution. Integrating DER requires significant attention to reactive power, making data on reactive power as valuable as that for real power. High time resolution is required to accurately assess network losses, given significant short-term changes in cloud cover that impacts solar customers and unbalanced phases, both common in LV networks. Even load data with 5- or 15-minute sampling resolution, as is available to Victorian DNSPs, is likely too coarse for reasonably accurate analysis.

Given the prevalence on WattWatchers, SwitchDin and other DER monitoring devices, the required datasets to synthesise realistic customer load profiles likely already exist. Such profiles would be quite useful both to researchers and DNSPs. This project revealed that DNSPs today often rely on load data for planning that is similarly inaccurate to that available to researchers.

A second opportunity for improvement is to develop novel libraries of cable and overhead lines based on first principles. Aligning industry on a common library of cables and overhead lines could significantly simplify data harmonisation, processing, and analysis, and prove invaluable in developing impedance data and building harmonic power flow models.

Digital Skills and Capability

Due to the large quantities of data in propriety data formats with a complex data model, the key challenges in this project were at the intersection of data science and software engineering, with input from domain expertise required for integration. Resourcing such multidisciplinary projects is well-known to be challenging, and CSIRO struggled to identify the appropriate skill sets required at the project's outset. At first, we put electrical engineers in charge of this project, whereas the more sensible approach would have been to have software engineers or data scientists take the lead. While eventually we identified the proper team, the project experienced growing pains along the way.

Anecdotal evidence from both other ARENA funded trials and our DNSP colleagues suggest CSIRO is not alone in struggling to simply identify the proper project management structures and logic for data-intensive electrical engineering research and innovation projects. Indeed, a US Department of Energy report[6] anonymously quoted a utility engineer, 'If you have an idea that you have a lot of model work to do and you're at the beginning of the project, you should probably think you've got five to 10 times the amount of work that you think you have,' the engineer said. 'That's only a slight exaggeration. It's an enormous undertaking to get real-time state estimation working on your whole network.'

[6] https://www.landisgyr.com.au/ezine-article/clean-data-models/. Readers may also find value in Voices of Experience: Insight into Advanced Distribution Management Systems (the headline report referenced above): https://www.energy.gov/sites/prod/files/2015/02/f19/Voices%20of%20Experience%20-%20Advanced%20Distribution%20Management%20Systems%20February%202015.pdf

Another challenge is that network data are graph-based datasets. In other words, some features of these datasets correspond to edge/line properties (e.g. lines and transformers) and some features correspond to vertex/node features (e.g. voltage limits). Processing such data requires understanding and maintaining this topological information, but many of today's data scientists are trained to work with time-series or tabular data. Most data science tools have this focus as well. In short, there is a dearth of appropriately trained and resourced data scientists to most effectively exploit the network data that is available.

There is a clear need to improve digital literacy within energy researchers, DNSPs, and the broader industry, including regulators. One approach to beginning to build such skills would be to establish an existing baseline of digital capabilities across relevant industry sectors, such as DNSPs, regulators and researchers, along with establishing several data related goals. With such information in-hand, a coherent plan to build the necessary skills and develop the appropriate tools could be established.

Data Sharing

Despite wide agreement on the need for greater data sharing among the electricity industry, researchers, and the public, doing so in an effective manner remains a challenge, and this project was not immune to such hurdles. There are a number of reasons for data holders to be hesitant about sharing, three of the most common are i) related to privacy and/or security concerns, ii) a lack of capability or resources to facilitate data sharing even if the data holder and recipient want to collaborate, and iii) a concern about the risk of losing or sacrificing some (often unquantified) commercial opportunity.

Data that can comprise sensitive customer information should be protected, and several approaches exist to effectively separate customer data from network data. These approaches can resolve many of the most immediate data sharing challenges – at least as they relate to sharing general network topology data. Indeed, planning engineers at DNSPs already often use synthetic or representative data for customers when constructing network designs.

One potential, albeit partial, solution would be to develop differential privacy-based network data cleaning tools (Fioretto, et al., 2020). Differential privacy can give mathematically provable guarantees that no information is leaked (Dwork & Roth, 2013), providing an approach that could be used to obfuscate sensitive data while maintaining representativeness of the electrical engineering features of the network. Continued research is necessary to extend this approach to distribution networks while avoiding pitfalls and misapplication (Domingo-Ferre, et al., 2020).

In general, significantly greater attention needs to be paid to data sharing – by researchers, DNSPs, and policymakers – in order to overcome the hurdles identified above. Australia's electricity DNSP businesses are granted a monopoly to provide an essential public service. As such, sharing their data with the broader public for legitimate uses – including developing new tools and methods to better plan and operate power networks – should be prioritised, while ensuring that reasonable privacy and security concerns are recognised and accommodated. As noted above, the Energy Security Board's data strategy explicitly calls for increased transparency of network data, and AEMC and AER are currently working on related initiatives. Improving data literacy within DNSPs and the broader industry – including regulators, researchers, and technology providers – is an

essential step in unlocking data to address these sharing challenges[7]. If nothing else, hopefully this project has helped highlight some of these data sharing problems and made a small but meaningful impact on the digital literacy of those engaged.

Potential Next Steps for Research

1. Establish benchmark network data sets for assessing the value of various technologies and models that attempt to estimate network values not directly measured. To this end, collect and publish extensive data on a relatively small number of real networks (approximately a dozen).
2. Establish typical, realistic, publicly available, customer load profiles to better represent the impact of DER customers on power quality by synthesising high time resolution (e.g. one-minute) data for both separate active and reactive power profiles.
3. Simplify network data standardisation, processing, and analysis, by developing novel libraries of cable and overhead lines based on first principles.
4. Develop a plan for building the industry's digital literacy skills by establishing an existing baseline of digital capabilities across relevant industry sectors, such as networks, regulators, and researchers.
5. Develop differential privacy-based network data cleaning tools to obfuscate sensitive data while maintaining representativeness of the electrical engineering features.

[7] https://www.oecd.org/publications/building-digital-workforce-capacity-and-skills-for-data-intensive-science-e08aa3bb-en.htm

Technical Summary

The Low-Voltage Feeder Taxonomy (LVFT) project had two primary aims. The first was to develop a realistic, publicly available dataset and models describing the most common types of LV networks found in Australia. The second was to improve understanding of the challenges and opportunities associated with using data to improve DER integration and to encourage greater overall network utilisation.

This report focuses primarily on the first objective; CSIRO partnered with several DNSPs to collect and analyse existing network data from more than 90 thousand low-voltage networks to identify an archetypal set of 23 low-voltage networks and build a number of associated models of these networks demonstrating the impact of DER operation on various network metrics.

The second primary project objective – better understanding the challenges and opportunities associated with network data – has proven to be a fruitful area for initial research. In short, there is a large gap between the data, tools, skills and capabilities required to enable a low-cost bidirectional grid, and that which currently exists in the sector. The key learnings and opportunities for improvement can be categorised into four separate topics: i) the role of network models; ii) improvements in data collection and cleaning; iii) digital skills and capability; and iv) data sharing.

To build the library describing the most common types of LV networks found in Australia, a data-driven identification of representative low-voltage networks was pursued, generated using tens of thousands of low-voltage network power flow models provided by DNSPs. However, there were two challenges observed about the data during this process. Firstly the locational distribution of network data was strongly biased towards south-eastern Australia, with over 90% of low-voltage networks in the dataset. Secondly, the varying levels of completeness of data among providers resulted in the working data being a reduced subset of the original, losing certain parts of impedance and load information. After scaling and dimensionality reduction were applied to the data, partition-based *k*-medoids clustering was chosen and the cluster hyperparameters were tuned. This resulted in a set of 23 representative low-voltage networks.

These 23 low-voltage network data sets have been disseminated for free and unfettered use by the energy sector via the publicly accessible data platform the Australian National Energy Analytics Research (NEAR, near.csiro.au)[8] platform. The network data sets are accompanied by notebooks that illustrate their use for power systems analysis, with OpenDSS as the simulation engine for unbalanced power flow analysis. In the notebooks, a user can set up simulation case studies with residential loads, PV systems, battery storage systems and voltage reduction demand response.

This set of networks and their associated models should provide useful tools to researchers and technology companies that require baseline data for testing different approaches for improving DER integration. The purpose of a testbed is to inform and facilitate early exploration in simulation

[8] NEAR was formerly known as the Energy Use Data Model (EUDM) https://research.csiro.au/distributed-systems-security/projects/energy-data-use-model/

Download via https://near.csiro.au/assets/f325fb3c-2dcd-410c-97a8-e55dc68b8064 .

related to the integration options of new energy technologies. In the absence of more accurate data on Australia's low-voltage networks, the enclosed set of network models is likely the best publicly available dataset.

1 Introduction

CSIRO, several Distribution Network Service Providers (DNSPs) and Energy Networks Australia (ENA) undertook the Low-Voltage Feeder Taxonomy project to help accelerate improvement of Australia's management of low-voltage distribution networks, including the transition to a bidirectional energy grid.

Currently in Australia, of particular interest for low-voltage (LV) networks is that some amount of rooftop solar is regularly curtailed because of network constraints, and every indication is that such curtailment will increasingly occur. Both DNSP operators and consumers would prefer for such consequences to be minimised. However, identifying and minimising low-voltage network constraints requires greater visibility of network conditions. Ideally this would be realised by access to power flow models of every given low-voltage network in which constraints are, or are likely to be, regularly occurring. Building such a comprehensive set of power flow models will take a considerable amount of time and additional resources, including significantly more data than anyone – including DNSPs – currently has.

The purpose of the LV Feeder Taxonomy project was to use the existing DNSP data to develop an initial set of power flow models for several representative low-voltage network types commonly found throughout Australia. These models would constitute key components of a desktop testbed for use by DNSPs, researchers, and technology providers.

This report is a primary deliverable of the project, along with the desktop testbed, and regular consultation with our DNSP partners. The project follows on from the Medium Voltage Feeder Taxonomy Project (Berry, et al., 2013), in which CSIRO developed a representative set of medium-voltage feeders for Australia. The taxonomy presented in this report comprises a set of *low-voltage networks* that represent those found in the networks of Australian DNSPs. It includes power flow models of these representative low-voltage networks, and models of distributed energy system components such as small PV generation and battery systems that could be potentially installed on such low-voltage networks. Together, the low-voltage network power models and distributed energy component models represent key elements of a desktop testbed for assessing the prospective performance of such low-voltage networks in the future under conditions of, for example, high levels of distributed energy technology capacity. The testbed can be used to examine various strategies for planning and operating a bidirectional grid with distributed solar, batteries, responsive devices and electric vehicles.

The taxonomy was developed via a data-driven identification of representative low-voltage networks based on tens of thousands of LV network models provided by DNSPs. The resulting models of 23 single transformer low-voltage networks (see box below) have been disseminated for free and unfettered use by the energy sector via the publicly accessible data platform the

Australian National Energy Analytics Research (NEAR, near.csiro.au)[9] platform, and advertised via webinars with DNSPs and other industry actors.

What is a 'low-voltage network'? A note on terminology.

The principal unit of analysis and clustering for this project is an individual 'single transformer low-voltage distribution network', which is defined as all components of a low-voltage (<1kV) power distribution network between an individual low-voltage distribution transformer and the customer power meters. For this project, it usually includes the distribution transformer. In principle this also includes connection service cables and customer points of connection, although whether or not the service cables are included is not critical for this project.

We use the term 'low-voltage network' (LVN), and sometimes simply 'network' as shorthand for 'single transformer low-voltage distribution network', even though the ordinary interpretation of a LVN (or network) could include other (larger or smaller) power network subsystems.

Feeder is a general term commonly used for both (material) transportation and (information or power) transmission networks to describe any link that is designed to connect a more centralised portion of the network which supplies distributed nodes in the direction of the product destination. There is no standardised definition that is sufficient to rigorously determine the precise boundaries of a feeder. In this project we have effectively defined a feeder as part of a radial 'low-voltage network' as a single (unbranched) link from the transformer power source to serve customers, excluding service lines. A 'low-voltage network' may include one or more feeders. A feeder may include one or more (electromagnetically coupled) conductors.

The power flow models are of generic representative low-voltage networks. As such, they are not directly suitable for detailed analysis of any specific low-voltage network as part of DNSP network planning. However, they can be used to develop an indicative assessment of broad planning and operational strategies to address network configurations that are expected to be typical for wide classes of networks. For example, these models could run scenarios on distributed energy resource (DER) adoption to see how often congestions are likely to occur on a network of a particular type with a given amount of DER. With an estimate of how many, and which types of, low-voltage networks are on a substation, one could extrapolate these models to better identify the costs and benefits of DERs for such a substation or for the whole of Australia.

One important learning from the project has been the sizable gap between the amount of data required to manage a bidirectional grid and the current amount of reliable data that is easily accessible. Improvements in data collection, sharing and analysis could meaningfully improve both planning and operations in Australia's low-voltage networks, and CSIRO and our partners

[9] NEAR was formerly known as the Energy Use Data Model (EUDM) https://research.csiro.au/distributed-systems-security/projects/energy-data-use-model/

Download via https://near.csiro.au/assets/f325fb3c-2dcd-410c-97a8-e55dc68b8064 .

anticipate several additional projects to help close the data gap and improve the sector's digital capabilities.

What is a 'low voltage'?
Generally, the cut-off voltage magnitude for what is 'low-voltage' distribution in public electricity networks is 1 kV RMS (root mean square). We note there are several key real-world configurations that are captured:

- single-phase, phase-to-neutral, typically between 230 V to 250 V RMS
- three-phase wye, typically 400 V to 433 V RMS line-to-line and 230 V to 250 V RMS line-to-neutral
- three-phase delta, typically 230 V to 250 V RMS line-to-line
- split-phase, typically 230 V to 250 V RMS line-to-neutral, 460 V to 500 V RMS line-to-line.

This report is structured as follows: the remainder of the introduction reviews related projects and literature. This is followed by a description of the analysis process that derives the final set of 23 network models. Chapter 2 describes the process of converting the network parameter data supplied by DNSPs into datasets suitable for performing clustering analysis, which is then described in Chapter 3. The results of the clustering analysis are presented in Chapter 4. Next, Chapter 5 discusses the base case power flow results, after adding time series data for loads. Moreover, Chapter 6 illustrates the use of those base cases across deployment scenarios for PV, batteries and demand response. Chapter 7 derives and discusses lessons learnt. Finally, Chapter 8 presents the conclusions and suggests options for future work.

1.1 Review of related projects

The Low-Voltage Feeder Taxonomy project sits within a wider context of similar national and international efforts to develop improved models of both electrical power networks and customer loads. These efforts are guided by an aim to enhance operational management of networks, with a particular contemporary interest in the management of distributed energy resources such as photovoltaic (PV) generation, batteries and electric vehicles (EVs), minimising energy losses while maintaining power quality. The Low-Voltage Feeder Taxonomy project follows on from Berry, et al. (2013) which produced a set of 19 medium voltage (11kV) feeders that are, in a sense, representative of those throughout Australia, with associated network models. These 19 representative medium voltage feeders were selected after statistical clustering analysis of 370 feeders from 11 Australian distribution network service providers.

The Low-Voltage Feeder Taxonomy project was conceived in the Australian context and complements other ARENA-funded research. Therefore, we reviewed a set of related projects in Australia, as well as international projects with similar methods and goals. We specifically noted a range of emerging topics in low-voltage grids: system identification & network data cleaning, state estimation (also known as low-voltage network visibility) and optimal control solutions such as distributed energy management systems (DERMS).

1.2 Projects that include identification of representative distribution networks

The Low-Voltage Feeder Taxonomy has identified a set of 23 low-voltage (415V) low-voltage networks after statistical clustering analysis based on 41 network features of network models. These were selected from more than 90 thousand low-voltage networks from 7 Australian distribution network service providers. It has produced and made available software models of these 23 reasonably representative networks and collated a set of time series models of load and PV generation that can be associated with the network models to undertake power flow modelling analysis. The load and generation time series, however, were not selected via a rigorously systematic statistical analysis process.

In contrast, Dale (2013) describes the selection of a set of ten distinct low-voltage network load temporal profiles, clustered according to the patterns of (real) power delivery. These profiles were selected from measurement of load in south Wales, Great Britain, from over 800 substations and 3600 voltage monitors. After being clustered based on substation load profiles (that is, not network parameters), the electrical network and customer characteristics of the clusters were inspected, resulting in a statistical model allowing a given low-voltage network to be classified as most likely belonging to one of the ten load profile clusters on the basis of network and customer characteristics alone. Each substation cluster was further divided into substations with or without one of five specific 'low carbon' technology interventions (e.g. PV, EV, heat pumps) in order to investigate whether these had a significant impact on the net load and voltage analysis based on monitored data rather than power flow modelling (Li & Shaddick, 2014; see also Li & Shaddick 2013).

In a similar project, Rigoni and Ochoa (2014) selected 11 low-voltage network feeders as representative of those of Electricity North West, a distribution network operator in the northwest of England. These were also selected using clustering statistical analysis, however the clustering was based on both static network characteristics and dynamically variable monitored data over a winter season, including not only voltages and loads but also temperatures and harmonics. Out of 523 feeders from 127 networks with network models, there were 383 feeders that also had monitored data available.

In addition to the 11 selected representative Electricity North West low-voltage feeders selected by Rigoni and Ochoa (2014), Navarro-Espinosa (2014) (see also Navarro-Espinosa 2016) describe the public release of models of a further 25 networks with a combined total of 131 feeders with enough detail to allow three-phase, four-wire power flow analysis. These detailed models were developed from (static) Geographic Information System (GIS) data from the network owners and validated using extensively monitored, dynamic, operational data. These network models were released in combination with synthesised time series of load and low carbon technologies. Using these 25 network models a Monte Carlo analysis was performed to assess the 'hosting capacity' for these networks in terms of penetration percentage uptake of low-carbon technologies that can be included before voltage or thermal power quality limits are reached (Electricity North West, 2014).

1.3 Related ARENA-funded projects in distribution networks

Another ARENA-funded project in a low-voltage distribution network is a PV hosting capacity analysis process for selected network models in Victoria, Australia (Procopiou, et al., 2020). Out of 351 medium voltage (11-22kV) feeders owned by a single distribution network service provider, four (4) representatives were selected and models obtained. The original set of 351 MV feeders was split into two categories, namely urban and rural (which were defined *a priori* rather than being data driven) with two representatives from each category selected based on a handful of customer and feeder physical features (Procopiou & Ochoa, 2019).

These 4 medium voltage models that are based on network operator data are combined with artificially synthesised models of low-voltage feeders. The low-voltage feeder models are not based on existing low-voltage feeder infrastructure with topologies and customer locations known in detail. Rather, the models are constructed using network company design standards for low-voltage feeders based on maximum demand at the feeder head, and parameters such as customer numbers, sectors and line lengths, which are estimated if not available. A set of time series of load and PV generation profiles was synthesised based on measurement data from 3000 customers with smart meters.

The primary focus of the project is power flow analysis for PV capacity studies rather than the development of a robust set of network models and associated load data. To this end, the medium and low-voltage network models were combined with the models of operational load and generation in order to understand the relationship between voltage stability performance and the quantity of PV capacity installed (Procopiou & Ochoa, 2019). The potential to improve performance of low-voltage networks with high PV penetration using traditional solutions such as tap changing transformers and network augmentation was reported in Procopiou, Pertrou and Ochoa (2020). The potential for further improvement again with more advanced PV controls and customer battery energy storage with and without sophisticated controls is reported in Procopiou, Liu and Nacmanson (2020).

In another ARENA project, Krause (2019) describes the development of state estimation technology. This enables the real time operational performance of an electrical network at numerous multiple locations to be estimated based on a small amount of measured data and knowledge of network parameters. This approach permits the management of electrical network performance with fewer direct measurements to, for example, increase permissible PV generation. However, it is difficult to empirically verify the accuracy of state estimation without the installation of costly monitoring equipment and to undertake the verification across a broad range of typical real-world installations.

The low-voltage feeder taxonomy could provide a valuable representative set of low-voltage network models and sample load time series data to verify in simulation and thereby improve technologies such as those described by Krause (2019) and to do so across a broadly range of low-voltage networks that are typical to the Australian context. Models made available by the low-voltage feeder taxonomy project could be used to simulate both the estimation accuracy of such technology and to quantify the improvement in network performance in terms of distribution technology hosting capacity that would be enabled by state estimation technology.

The Low-Voltage Feeder Taxonomy project could also potentially enhance projects such as Dinning, et al. (2020), which conducted simulation studies to estimate the PV hosting capacity of another Victorian distribution network service provider organisation (Citipower and Powercor: CPPAL). The conclusions of this report were based on 10 examples of models of low-voltage network and although those ten were intended to capture a wide variety of network types, one limitation of the analysis was the challenge in confidently extrapolating results concluded for such a sample as applicable across the broader network. Dinning, et al. (2020) also point out that, due to limitations in the data provided to them, their analysis was undertaken using network models under assumptions of a balanced three-phase load. This approach leads to optimistic assessments of modelled performance relative to empirical measurements from advanced metering infrastructure. The network models provided by the LVFT are unbalanced three-phase four-wire and support the more accurate unbalanced power flow analysis.

1.4 Review of scientific literature

Related literature has been reviewed where network test models have been proposed as a common benchmark or where clustering analysis has been used with models of electrical networks to simplify a much larger set.

Identified network models in the literature are from numerous geographical locations in Europe and North America, as well as a few from Australia, and include both low-voltage and medium voltage networks. However, there is little standardisation in network data file format and/or power flow analysis software used. Across the literature reviewed, there also appears to be limited standardisation regarding the appropriate unit of analysis for clustering – i.e., what is the portion of electrical network that is the appropriate scale for analysis or recombination. The literature surveyed provides broadly general support for the importance of having publicly available sets of representative network models for the purposes of testing. However, the Australian studies identified provided models only for medium voltage networks or were limited in scope to only one or two DNSPs.

The literature demonstrates that clustering analysis appears to be a suitable method for selecting representative examples of large sets of electrical network subsystems and load profiles, even though there is limited analysis of the most appropriate classification relevant features for alternative applications. Applications of test feeder models that are of particular interest in more recent publications include understanding the potential impact of distributed energy resources on the performance of electrical power networks.

Marcos, et al. (2017) presents a review of publicly available test feeder models and model sets, with a focus on models of United States feeders. They point out that aside from modelling an actual feeder, selected for desirable characteristics, it is alternatively possible to develop a test feeder model by constructing a synthetic network. This can be accomplished by composing together sections of models of actual networks that have been identified by clustering as being similar to each other. Alternatively, synthetic networks can be designed manually to exhibit particularly desired features, or by using planning and network design tools and processes used in practice by distribution planners. Although Marcos, et al. (2017) identifies numerous publicly available test feeder models, it finds that few are intended to be representative, identifying only two such sets. One is a set of 12 representative ‘prototypical feeder models’ selected from 27000

Pacific Gas and Electric feeders. A second is a taxonomy of 24 prototypical radial distribution feeder models discussed further below (Schneider, et al., 2009). Both of these publicly available model sets used automated clustering methods to identify representative feeders and are made up of models of actual feeders.

There have been numerous studies that have used clustering analysis to develop a smaller representative set of network models from a much larger set (see Table 1 and Table 2 below. See also Table 2 of Ma et al. 2019 and Ma, 2020 for further examples of network model clustering). They are generally motivated by the idea that detailed analysis of a representative set, rather than the full set, is a lower cost alternative that still allows relatively valid conclusions to be drawn about the full population.

Many of these studies are primarily motivated by PV hosting capacity analysis in particular or more general evaluation of smart grid technologies, although other types of intended application include general network planning, grid loss modelling and reliability analysis. It is worth noting that highly detailed network models and individual customer load profiles may not always be required for studying the impact of PV generation on network voltage performance, as a two-bus simplified equivalent model may be sufficient for some limited purposes (Santos-Martin & Lemon, 2016, see Section C in particular). More generic methods for distribution feeder model simplification, while maintaining simulation accuracy, appear in Pecenak, Disfani, Reno, & Kleissl (2018).

Typical sizes of the initial model set range from hundreds to tens of thousands, which are reduced to representative sets of tens to hundreds. Earlier studies tend to rely a little more on expert analysis and non-automated decision making in the processes of identifying useful network characteristics to be used as a basis for classification and in defining initial categories within which further clustering disaggregation is to take place. Earlier studies also tend to rely more on traditional statistical techniques such as principal component analysis (PCA) and analysis of variance (ANOVA) with later studies exploiting more automated and nonlinear analysis techniques such as automated hierarchical classification and *k*-means clustering and its variants.

For most of the papers identified, the primary focus is on the results of analysis of the representative set, as opposed to the development of the representative set *per se*. Exceptions to this general observation include Schneider, et al., (2009) (see detailed technical report in Schneider et al., (2008)), Berry et al., (2013), and to some extent Rigoni and Ochoa (2014) and Li and Wolfs (2014). Schneider, et al., (2009) originally made the network models publicly available (in 'GridLAB-D' format). Berry, et al., (2013) also originally released the network models publicly (in Sincal format) through the Ausgrid Data Clearing House, and included indicative load profiles. It is now located at the CSIRO Data Access Portal[10].

[10] Representative Australian Electricity Feeders with load and solar generation profiles https://doi.org/10.4225/08/5631B1DF6F1A0

Table 1 Literature on Clustered Electrical network models - key features

	Application	Location	Sample Size		Data	Initial Feature Set	Initial Feature Set
Ma, et al. (2019)	Grid Loss modelling	Germany, 90 towns in Bavaria	5000 grids 31000 feeders	LV models	Automated model creation from GIS data		Includes load models
Shafiei, et al. (2019)	PV hosting capacity	Australian (likely Qld) MV distribution network	500 nodes	MV	Outliers defined as 7 SD on MV node voltages		Time based voltage magnitude only
Jain & Mather (2018)	PV hosting capacity	North American utility	3000	Distribution feeders		30	Includes Aggregate load data
EA Technology Ltd (2018)	LV network management (DER)	South Australia	1270 HV 56862LV	HV and LV networks			Geography, Peak load, customer count, construction-
Nijhuis, Gibescu, & Cobben (2015)	Assessing the future loading for network planning	Liander, the Netherlands	88000	LV feeders			Network and some load
Rigoni (2016)	PV hosting capacity	UK, 131 networks	232	LV feeders	383 cleaned to 232		network and customer X-istics
Li & Wolfs (2014)	Evaluating smart grid technologies	Western Australia	204	MV (22kV) feeders,		34 MV features	Network, customer and load
Li & Wolfs (2014)			8858	LV feeders		26 (LV)	
Berry, et al. (2013)	Evaluating smart grid technologies	11 out of 16 Australian DNSPs	370	MV feeders			Network parameters and customer scale.
Broderick & Williams (2013)	Screening for PV hosting capacity	California	3000	Distribution feeders		15	Network and load data
Dickert, Domagk, & Schegner (2013)	Feasibility?	German		Low-voltage distribution networks		6 numerical variables	Supply obligation, network properties. More features would have been nice to have.
Schneider, et al. (2009)	Analysis of new smart grid technologies	17 continental US utilities and 151 substations	575	Distribution feeders, 12.75-35kV	GridLabD from SynerGEE	~35	Network parameters and customer scale.
Levi, Strbac, & Allan (2005)	Reliability , alternative investment strategies	Not identified	100 for the case study	Urban (i.e. underground) feeders	Excel	10 categories	'Structural' and 'population' categories (not load)

Table 2 Literature on Clustered Electrical network models - clustering parameters

	Feature Selection	Final Features	Clustering method	Final Cluster Set	Loads for Analysis	Load Data Link	Model data release
Ma, et al. (2019)		8	SPCA	300	Normalised standardised load profiles	https://www.bdew.de/energie/standardlastprofile-strom/	Not found
Shafiei et al. (2019)	Initially clustered into 5 groups		2D *k*-Means abs(V) time	4-5 per group	Not directly applicable		Not found
Jain & Mather (2018)	Covariance heat map: Kendall rank correlation coefficients: Principal component analysis;	19 after reduction	*k*-medoids, Partitioning Around Medoids	11 finalised	PV profile source not specified		Not found
EA Technology Ltd (2018)	Expert selection	7 for classification Representatives based on different features	Based on expert selected features	7 HV networks 15 LV networks	6 representative days for 'typical' load profiles of 'different types' of customer		Not found
Nijhuis, Gibescu, & Cobben (2015)	Expert selection	About 15	Fuzzy *k*-medians	94 classes	15-minute load profiles from Energy Data Services Netherlands, (online) Nov 2014	https://www.kaggle.com/lucabasa/dutch-energy	Not found
Rigoni (2016)	With and without DG separated		hierarchical clustering, - ++, improved - ++, and Gaussian Mixture—GMM	11	Elexon Standardised Customer Load profiles	https://www.elexon.co.uk/operations-settlement/profiling/	
Li & Wolfs (2014)		6 variables incl. annual load	Ward's hierarchical cluster & discriminant analysis	9	Future Work		
Li & Wolfs (2014)		7		8			
Berry, et al. (2013)	Expert selection and cluster testing	14	*k*-medoids	19	Feeder-head SCADA data from principally resi, comm, or ind feeders to build class-based load profiles.	'illustrative' profiles released	Yes, but maybe no longer available
Broderick & Williams (2013)	Correlation analysis	11 features	*k*-means Cubic Clustering Criterion	20 feeders	Future Work		
Dickert, Domagk, & Schegner (2013)		2 principal components	PCA, k-means	6 groups, 3 reps (good, average, worst) each group	Companion paper for load model research		
Schneider, et al. (2009)	Initially clustered into categories based on climate zone and voltage	35	ANOVA	23, 1 from each category + 1	Unclear		https://www.gridlabd.org/ see Schneider, et al., 2008
Levi, Strbac, & Allan (2005)	Appears expert selection	4 features selected	Appears expert driven hierarchical	19 for case study demonstration	Fault analysis, not power flow		Not found

Most of the initial features identified for the purpose of the cluster classification include not only network characteristics, but also customer characteristics, including (time-aggregated) load data. A few studies use time-varying load profiles for analysis of the behaviour of the representative networks subsequent to their identification. These include a couple of studies using publicly available load profiles standardised as representative of particular customer categories (Ma, et al., 2019; Rigoni & Ochoa, 2014). One study undertook analysis using much more diversified load profiles at fifteen-minute intervals (also publicly available from a Netherlands website) (Nijhuis, Gibescu, & Cobben, 2015).

Most papers used cluster metrics that respect the initially provided feature variables. In contrast, the clustering method of Ma et al. (2019), 'supervised principal component analysis', is designed to find clusters that are specifically relevant to a particular purpose, weighting features that best explain the target variable, in this case, grid losses.

The importance of, and difficulty of producing, reliably good models incorporating reliable data is mentioned occasionally, although data cleaning methods tend not to be described in detail. (Schneider, et al., 2009) asserts that 'the issue of data quality and data consistency was one of the largest issues that had to be addressed' and Jain and Mather (2018) also emphasises that data quality checking is an important activity that 'must be completed before the cluster analysis is used' because it is 'highly susceptible to outliers, missing data, or bad data'. Ma et al. (2019) describe an automated method for developing the initial set of grid models, extracting distribution network data from a GIS database, and making inferences about electrical connectivity based on spatial proximity.

The power flow network models in Rigoni et al. (2016), Navarro-Espinosa and Ochoa (2016) and Procopiou and Ochoa (2019) support unbalanced three-phase power flow analysis. Ma et al. (2019) specifically mention that they perform only a balanced three-phase power flow analysis due to the limitations of the power flow analysis package and the lack of availability of data necessary for creating the appropriate models.

2 Data Set Ingestion and Feature Extraction

This section covers the description of the original power flow network datasets from the DNSPs, the extraction of the power flow model parameters and the checking and creation of selected features.

2.1 Data cleaning and transformations

The network data provided by seven DNSPs (out of 16 nationally) was condensed in the following file formats:

- Ausgrid: excel spreadsheets of a proprietary specification
- Ausnet, Essential, SAPN and TasNetworks: Sincal files backed by databases
- Endeavour, Energy Queensland: Powerfactory files.

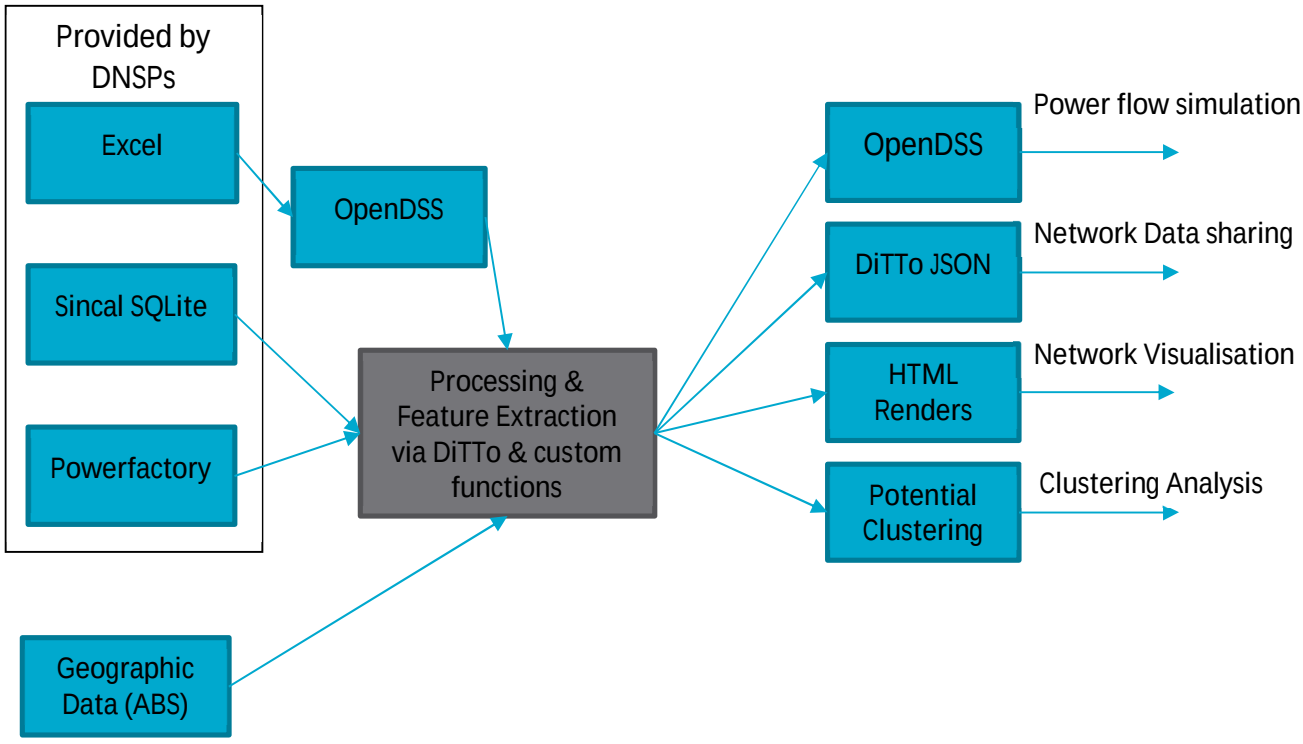


Figure 1 Simplified data flow diagram of data ingestion, processing and outputs

This section describes the processes and techniques used to clean and transform this network data into useful formats for enabling data-driven clustering, as described in the subsequent section. Figure 1 shows the high-level data flows from the ingestion and pre-processing steps. The network data from DNSPs was first converted to OpenDSS, an open data model, format. Next, the low-voltage parts of the networks were extracted (the medium voltage parts discarded) and split into smaller low-voltage networks, each with a single MV/LV transformer. Where necessary to ensure that the resulting topology of the low-voltage networks is radial (that is, without loops, known as 'cycles') some low-voltage links were removed. Individual feeders within each low-voltage network were identified to assist with visualisation. Finally, selected features of each low-

voltage network were calculated, based on the OpenDSS data model of each low-voltage network that had been split out from the original DNSP data files. A small number of features were determined based on location data associated with each low-voltage network and locational characteristics sourced primarily from the Australian Bureau of Statistics (ABS). The resulting network models and features were converted into various formats for further downstream processing. The following sections describe these steps in detail.

2.1.1 Data Ingestion

NREL's open source *DiTTo* Python framework (https://github.com/NREL/ditto) was used as the main data processing and conversion platform for reading these files into a common data model for visualisation, and feature extraction and normalisation, before outputting all LV networks to individual OpenDSS files which were used for power flow simulations.

Ausgrid spreadsheets were converted to OpenDSS using a custom-written function (in the Julia programming language) and read into DiTTo using its OpenDSS reader. Custom PowerFactory and PSS/Sincal parsers were added to DiTTo to ingest these additional file formats. Functionality to split a single medium-voltage network into multiple low-voltage networks was also added to these parsers. Splitting was done by identifying all MV-to-LV transformers in each file, then extracting all lines and nodes on the LV side of each transformer by systematic search, ensuring none were duplicated even if observed in subsequently processed networks.

References to 'network' in the following section generally refers to these separated low-voltage networks, unless otherwise specified. All feature extraction, clustering and simulations below are performed on such individual LV networks. Because of the lack of verifiable data on the subset of these networks referred to as 'feeders', feeder features could not be used for the following analysis. An attempt was made to identify feeders heuristically, but apart from appearing as bold lines in visualisations was ultimately not used for any analysis.

The final input dataset amounted to 5.2Gb of data in 330 files, and took about 12 hours to parse, split, visualise, extract features and convert to OpenDSS format. This resulted in approximately 94,700 individual low-voltage networks (see Table 3).

Table 3 LV network count by DNSP

DNSP	File Count	LV Network count
TasNetworks	310	27020
Essential	1	1017
SAPN	1	4723
Ausnet	1	46
Ausgrid	10	60182
Energy Queensland	2	34
Endeavour	5	1633
Total	**330**	**94655**

2.1.2 Ancillary data and visualisation

Ancillary data was added by looking up Meshblock and Remoteness Area categories based on the mean latitude/longitude of all provided coordinates for each LV network as some crossed multiple category boundaries. This ancillary data is sourced from the Australian Bureau of Statistics (ABS). Meshblocks[11] are the smallest geographical area used by the ABS, they broadly identify land use such as residential, commercial, primary production, etc. Remoteness Area[12] structures divide Australia into five objective classes or remoteness based on relative access to services, categories include: Major Cities, Inner Regional, Outer Regional, Remote and Very Remote.

A visualisation package was added to DiTTo that used NetworkX (https://networkx.org/) and Pyvis (https://pyvis.readthedocs.io) to render each network to a Javascript based force-directed graph visualisation that can be redistributed as a Hypertext Markup Language (HTML) file. See Figure 2 for an example rendering.

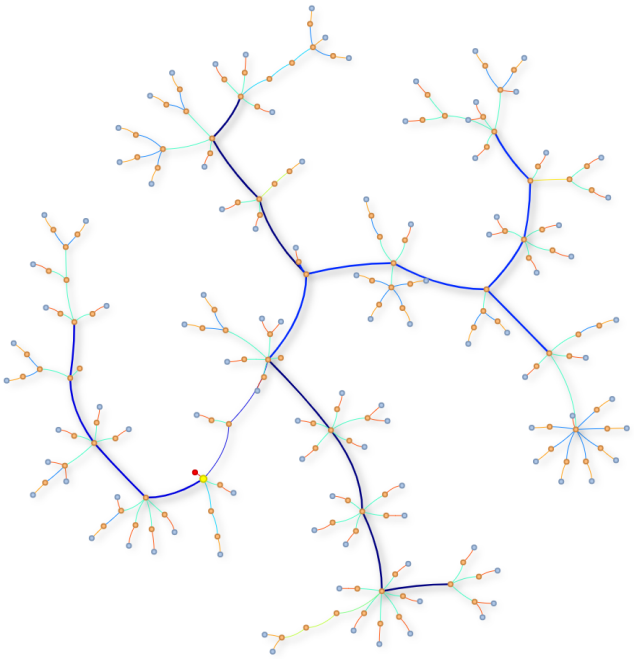


Figure 2 An example LV network visualisation

These visualisations were used for initial validation of the parsed network structures, helping to identify and correct early issues with

- disconnected network segments
- cycles/loops in some networks caused by closed switches
- inconsistent power-source/transformer ordering.

---

[11] ABS Meshblock 2016 data is sourced from : https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1270.0.55.001~July%202016~Main%20Features~Mesh%20Blocks%20(MB)~10012

[12] ABS Remoteness Area 2016 data is sourced from https://www.abs.gov.au/websitedbs/d3310114.nsf/home/remoteness+structure

2.1.3 Cycle removal

A significant number of LV network models contained cycles (that is, were meshed not radial) after parsing. Cycles can be easily identified using established graph processing algorithms once the ditto models are constructed. We observed that almost all cycles were caused by switch-type network elements (fuses, breakers, sectionalisers, etc) which were marked as closed in the supplied network model. In the absence of other information about whether these modelled loops were physically accurate, we assumed that they should be removed as this would otherwise be a very unusual (though not impossible) configuration.

To select which links to remove, each *cycle basis* (the minimal set of simple cycles that can be selectively summed to form all possible graph cycles) in a network's graph was traversed in a transformer-first direction, and the first *trivial* line (defined in 2.1.4) removed. This is broadly equivalent to opening the first switch discovered in the cycle. Trivial lines were regarded as candidate switches, as the source data did not always mark such lines consistently, however network visualisations showed they were usually identifiable in a handful of other ways (e.g. short lengths, missing impedance, certain string identifiers). The removal of suspected switches was repeated until no cycles remained in the network model, or no further trivial lines which could be removed remained.

A simple automated process could remove cycles from around 3000 LV networks, leaving cycles in only about 20 LV networks (which were later removed from the clustering dataset).

2.1.4 Feeder identification

As there was no explicit information on which segments of the LV network were identified by DNSPs as individual feeders, an automatable heuristic was developed to do so. The initial intention was to subsequently calculate feature data from feeder and non-feeder segments of each LV network for use in the clustering process. This feeder identification heuristic was designed to work with Low-Voltage Feeder Taxonomy network data, which in general lacks Line/Node labels that might have made feeder identification easier. It is general enough that it should function on arbitrary electrical networks, though this has not been tested.

Separating LV feeders from larger integrated network data sets is a two-step process. First, we find the low-voltage portions of the networks. That can be done in a fairly straightforward manner, by inspecting the voltage level data of the buses, or by identifying MV to LV transformers, and collecting the networks downstream of them. However, a single low-voltage network can contain multiple feeders. Feeders are the backbone of the distribution network, which connect to customers through service lines. The data we received does not explicitly identify individual feeders. Therefore, in order to separate them within a final step, a heuristic would need to be developed to identify individual feeders. However, this makes the simulation data less interesting from the electrical engineering perspective, so we decided to not separate the low-voltage networks into feeders. Feeders automatically identified by heuristic were later spot-checked by a domain expert, but not exhaustively verified. Consequently, feeder metrics were ultimately not used for clustering, though they are visible as thicker lines in the network renders to aid interpretation.

The heuristic used the following logic:

1. Identify the assumed *feeder head node* as the first degree >2 node (i.e. nodes with at least 2 connected Lines, ignoring those attached to loads) downstream from the low-voltage transformer.
    a. If step 1 fails it is usually because the feeder has no branches, in which case assume the first node down from the substation is the *feeder head node*.
2. Identify all the *trivial* lines and line-types. These are any with:
    a. low or missing (positive sequence) resistivity values
    b. a line type that is a switch, fuse or breaker (or with a name indicating such)
    c. very short or missing length values.
3. Identify all distinct line types in the network based on uniqueness of a set of features, including: positive sequence impedance, line type, nominal voltage, and line class.
4. For each distinct line type from step 3, take the subset of nodes/lines containing only that line-type and trivial lines (because switches/fuses can form part of a feeder).
5. *Connected-feeders* are then identified as any line-type subgraphs from step 4 which are connected to the *feeder head node* and have a *diameter* (in metres) greater than 10% of the whole network diameter[13] (diameter is the longest path across the network that does not traverse and node more than once).
6. We also define *remote feeders*, which are subgraphs that are not directly connected to the feeder head node but have a diameter greater than their distance to it.

2.1.5 Feature extraction

Once all source networks were parsed and converted to separate low-voltage DiTTo models, a set of about 200 metrics (features) was extracted from, or calculated for, each LV network. See **Appendix A Network Feature Descriptions** for a full list of metrics and descriptions. Broad categories of features included:

- geographical information, such as land-use and remoteness categories
- simple counts of each network element, and by voltage (240, 400 and >400V)
- statistics (min, max, mean, median) of cumulative physical and electrical characteristics of the path from each node to their LV transformer. For example, the average distance in metres and average impedance between all nodes and their transformer.
- metrics calculated by DiTTo's built-in functions (see https://nrel.github.io/ditto/metrics/).

Data limitations meant that some metrics were unreliable or could not be calculated. The metrics extracted using custom-written code for this project and those already produced by DiTTo often

[13] https://mathworld.wolfram.com/GraphDiameter.html#:~:text=The%20graph%20diameter%20of%20a,is%20a%20graph%20distance.

overlapped, providing redundant information. This redundancy was intentional; it was allowed so that no important metrics were missed if sufficient existing data permitted. Dimensionality reduction techniques used prior to clustering were later used to remove any such redundancy, so that no repeated metric unfairly weighted the cluster results.

2.1.6 Outputs from data ingestion and processing

The outputs from data ingestion and processing were:

- DiTTo files – LV network models were serialised to both JavaScript Object Notation (JSON, text) and Pickle (a Python object serialisation binary) files
- Comma Separated Variable (CSV) features – a single CSV table containing the final set of per-network features used for clustering
- Hypertext Markup Language (HTML) renders – individual visualisations of each LV network
- OpenDSS files by individual LV network– DiTTo models were converted to OpenDSS files; this conversion was not entirely reliable and some manual correction was necessary due to errors in the source data or conversion process which regularly prevented their direct use for power flow simulation.
- Python code – this included re-usable classes for plotting and feeder identification heuristics, plus PSS/Sincal and PowerFactory parsers.

2.2 A note on data limitations

If clustering is to be based on features that include power flow properties, then all the LV networks in the population to be clustered must be represented in a format suitable for power flow analysis. We were unable to achieve this breakthrough during the project. Processing the network model data proved to be a major proportion of the effort for this project. The data that was initially provided was required to be converted into:

- a database that is suitable for performing a reproducible clustering analysis that is sensitive to low-voltage network features of relevance to network performance, for each low-voltage network separately identified, and;
- a format that is suitable for power flow analysis, for at least cluster representative low-voltage networks (cluster medoids).

Many issues with data ingestion and feature extraction were ultimately caused by the lack of standardisation with the ways DNSPs describe and maintain network models, and limitations and proprietary nature of the software packages used to do so. There was too much data to be processed manually, and the availability of software tools to parse these proprietary file formats is lacking, creating significant barriers to their external use.

Varying levels of network data completeness were prevalent features of the majority of network models provided. This included

- **missing:** distributed generators, loads and service lines, grounding data, feeder information, phase labels for customers

- **inconsistent or missing:** switch labelling, impedance representations, and transformer configurations.
- **undocumented:** GIS coordinate systems

Even meeting the minimum data conversion requirements proved to be technically challenging because the data for the network models were:

- provided in several alternative original formats (including some proprietary)
- protected by privacy and intellectual property access restrictions (load data is particularly sensitive)
- partially incomplete (e.g. not all LV networks have been digitised into PowerFactory/ Sincal)

Further details of the data extraction challenges follow.

2.2.1 Format

Each model software format had its own requirement for accessing the data and its own data model structure for the network parameters of interest. The spreadsheet data could be extracted using readily available software libraries for reading Excel files. PowerFactory data was only accessible through its API, so a Powerfactory software license was needed for performing, developing, and testing the data extraction process. For the models in Sincal format, it was possible to read data directly from the SQLite or *mdb* (Microsoft Access Database) formats, enabling the data to be extracted without executing PSS/Sincal software. Network data in Sincal format was compatible with different PSS/Sincal software versions and the corresponding databases had different structures.

Because the software formats (and versions) had their own data structures, they each had their own challenges, which had to be solved independently. Even automated data extraction for PowerFactory proved to be slow, as large numbers of PowerFactory application programming interface (API) calls were necessary, each was quite slow, and they could only be single-threaded.

Developing data extraction software that supports the large variety of components that can be elements of a network model is challenging, as each component type, such as fuses, switches, transformers, or lines, has its own extraction and data transformation challenges:

- Component labels may be similar, but not identical, across data sources– for example: nodes and buses, switches and fuses, lines and branches.
- Specific parameters can be provided in nondimensionalised (per-unit) or absolute (SI) units.
- Impedance data can be instantiated in various ways, e.g. absolute impedance, length-normalised impedance, geometries for overhead lines and cables, sequence components vs phase impedances, etc.
- Some data models require three-phase loads to be represented with individual set points per phase as 3 single-phase loads; similar problems exist with three-phase transformers with independent taps or regulators.
- Some data was entered in non-standard ways even though the software supported standardisation. For example, switchable components (e.g. fuses, breakers) were often found (even when the source model provided explicit flags) to be represented by a generic

line with one or more of: a) a low or missing (positive sequence) resistivity value, b) a name containing: ‘removable’, ‘fuse’, ‘switch’ or ‘connector’, or c) a very short or missing length value. This meant that we had to rely on hand-checked heuristics to identify these components and schemes to selectively open-circuit them to remove cycles on the network graphs.

For reliable data extraction, support for specific components should be engineered for, not merely implemented on a trial-and-error basis as determined by data that is presented.

Some component types are not clearly represented in all formats. For example, a split-phase transformer is not necessarily labelled distinctly from a three-phase transformer but is indicated by the number of transformer parameters that have defined values.

Although PowerFactory supports explicit representation of 4-wire conductors, we rarely saw it being used in the data provided. When extracting sequence impedance data, it is therefore unclear whether this represents an actual 3-wire conductor, or is instead a 4-wire Kron-reduced approximate representation of an actual 4-wire conductor. In principle, additional information is needed to disambiguate.

This reinforces the importance of data model specification and documentation for both the source data to be extracted and the target data format to promote swift and error-free data extraction and conversion code.

Because of different data models, each format supported the representation of data parameters that other formats did not.

Further discrepancies between networks from different DNSPs were found, even those using the same software, these included:

- GIS Coordinates:
    - different coordinate systems (lat/long, WGS84, X/Y)
    - different coordinate densities: per-node, per-line, per-network
- different property sets between different file formats
    - missing/incomplete line impedances, types, wire diameters
- inconsistent/non-unique sub/node/load naming conventions

2.2.2 Confidentiality

All network data was treated as confidential. Only individuals within the CSIRO project team had access to the data, and even within the team data access was isolated to only what was necessary to perform a task. The data is sensitive, as it potentially contains electricity consumer identifying details and time series of power consumption by consumers. Furthermore, the original data contains physical coordinates, which can be cross-linked with consumers. Therefore, we removed those features from the public data release.

Due to confidentiality and privacy constraints, the software developer responsible for developing the data extraction code was not permitted direct access to the model data for actual networks. For development testing and debugging, they were provided a relatively small set of public test data. It eventuated that, for the purposes of testing, the test data was not representative of the actual data and its deficiencies. Testing the extraction code on the data for actual networks

required the involvement of a second person, who was permitted access to that data, in addition to the original software developer. This significantly complicated the development process, which was required to address numerous features in the data of the actual networks, but which had not been present in the test data set.

It follows that the developer writing the data extraction code must have direct access to the actual target data to enable a reasonable level of code robustness. Even setting up an appropriate test environment with appropriate access to all the required software licences proved administratively and logistically non-trivial.

Since load data is privacy sensitive, we were not able to use it. In order to perform load flow analysis, therefore, it became necessary to assign load data from public sources. This is a quite common strategy – in none of the related projects and network clustering papers inspected as part of the literature review did we find an example in which network data and matching load data were both present. Of the projects and papers where load data was used for power flow analysis, two used synthetic profiles intended to represent a class of customers, and one used profiles that were partially aggregated to preserve privacy, as well as being anonymised.

2.2.3 Data completeness, accuracy and bias

Various optional data elements were invariably missing from some data sets. Geographic coordinates, for example, were provided to varying degrees of completeness. Not every element is associated with defined geographic coordinates, coordinate components were sometimes inconsistent, or represented in various coordinate systems – including relative co-ordinates with no clear datum.

For most network models, line models contained physical path layout information, most of which is superfluous to electrical power flow analysis. This was generally because the physical pole and wire locations from GIS systems were used directly to model nodes and lines. This results in network models with numerous nodes that are not only redundant, producing unnecessarily large and complex representations, but can also lead to avoidable numerical challenges for power flow analysis algorithms. Other anomalies included branches or lines of zero length, which tend to break power flow solvers that don't filter for them, and therefore have to be eliminated through data transformation.

Network model data in different formats was also combined with information from other sources, which had not necessarily been cross-correlated for consistency. For example, latitude and longitude information on networks and network elements provided by the DNSPs was used to look up the likely land-use associated with each LV network using publicly available land-use data.

Many network models included line connectivity information but lacked phase connectivity information. However, both are necessary to unambiguously define an electrical topology. For many other network models, some line data was further incomplete or inaccurate. This includes missing impedance data, requiring inference from cable type, missing cable type specific impedance data, requiring inference from cable physical dimension specifications. It included missing mutual impedance between phases, and representation of impedance in a symmetrical component format that is incomplete (e.g. missing the zero-sequence component) or an

inaccurate approximation (e.g. a symmetrical three phase component approximation of a single phase load).

As it is highly desirable that the data for the features selected for clustering be adequately complete and consistent across the full population of low-voltage networks to be clustered, the problem of missing and suspect data in some of the network models created a dilemma. For some otherwise promisingly useful features which were missing or dubiously accurate from some low-voltage networks, it was necessary to decide whether to eliminate the features or the low-voltage networks from the clustering process, or whether to include them, and therefore risk biasing the clustering results differently.

The size of datasets was significantly different across DNSPs. There was as many as 60 thousand, compared to as few as 17, LV networks for different DNSPs. For all but two DNSPs, the provided network data were incomplete and were obviously biased samples of the actual LV network assets digitised to date by the DNSPs. As a result, these data issues undoubtedly biased results towards the more complete datasets, and likely resulted in the identified clusters missing important network topologies that are significantly less prevalent in the original dataset than in the field.

Owing to missing data, and data that was inaccurate or possibly inaccurate, it was required to check the data for adequate completeness and accuracy, making corrections where feasible. Manual cleaning, even manual inspection, of data is very time-consuming. It is not practical beyond tens of LV networks and certainly does not scale to the thousands or more instances in the source data.

2.2.4 Impedance data

Representing the physics of phase unbalance is generally considered appropriate practice in power distribution network analysis (Kersting & Dugan, 2006). Phase unbalance is caused by unbalanced, e.g. single-phase, loads as well as unequal conductor resistance and/or reactance. 'Unbalanced power flow' is the simulation technology used for distribution network analysis, and can identify network congestions of different kinds: transformer, line and cable overcurrent, over- and under-voltage and more. When developing unbalanced power flow studies of real-world distribution grids, it is important to use the appropriate simulation engines (i.e. unbalanced power flow) with high-quality data, to represent the physics of underground cables and overhead lines in a way that mirrors reality.

Resistance is an important property of lines and cables, itself a function of material type and cross section, but it is not the only one. Mutual impedances, capturing the inductive effects of one conductor on the others and vice versa, are key to model steady state voltages accurately. Multiconductor transmission line equations are required to accurately model the physics of phase unbalance, which is caused by unequal mutual inductance and/or by unbalanced loading.

The multiconductor transmission line equations represent the physics of a line as a 'Pi-section', i.e. shunt admittances on the ends of the line, and a series impedance.[14] For multiconductor systems, these impedance parameters are square matrices, where the size is determined by the number of

[14] For example, Kundur (1944), *Power System Stability and Control*, McGraw-Hill Education

conductors. We obtain values for those impedance matrices by solving Carson's equations or by finite element electromagnetic simulation. Carson's equations define self and mutual impedance values for the primitive circuit.

Now we are ready to solve a power flow as-is, or we can perform additional approximations. In the context of four-wire networks with multiple or sporadic grounding of the neutral, Kron's reduction is commonly used. Kron's reduction is applied under the assumption that the voltage in the neutral is close to zero. In those circumstances, the 4x4 impedance matrix can be transformed into an equivalent 3x3 one.

Common problems with impedance data included:

- missing service lines (the lines between the feeder and the power meter for billing)
- phase connectivity across laterals and service lines that are missing or randomised
- impedance values that are unreliable (e.g. resistance to inductance ratios for the sequence components that appear to be inconsistent)
- Kron reduced data for 4-wire sections, without access to non-Kron reduced form
- lack of neutral grounding information, except at transformers
- low diversity in transformer configurations.

In the data we received from the DNSPs, it was observed that virtually all LV network data was in Kron-reduced form. Furthermore, the impedance data was specified in sequence components (symmetrical components), i.e. the positive and zero sequence. Using such sequence representations is equivalent to assuming the network has been perfectly transposed, i.e. the conductors all have the same self-impedance value, as well as mutual impedance value.

Sequence components have proven useful in the description of networks that have regular transposition of conductors, as well as sufficient aggregation to assume the load is balanced. Therefore, in high- and medium-voltage system simulation, the approximation is usually satisfactory. Nevertheless, transposition does not occur in the LV system, and would not resolve the fact that some parts of the network are inherently unbalanced, i.e. single-phase branches. Furthermore, single-phase loads and DER are common, and the aggregation is not yet sufficient to cancel out.

2.2.5 Software development practices

The development of data processing workflows and software tools to implement them reinforced the importance of numerous basic software development standard practices. Ensuring that implemented algorithms perform reproducibly (at least during testing) is important for debugging. Developing unit tests is useful for detecting and correcting functional regression during debugging. Software version control during development is important, so automated version control tools can be helpful. With potential users of the data processing software running it on various operating systems within various software ecosystems, support for multi-platform operation, which requires multi-platform testing, is also desirable.

Where software development builds on an existing code base, it is necessary to judge their existing capability and maturity to assess their suitability for the task at hand, the risks of building on them, and the likelihood that the resulting product will be maintained.

Validation of the data extraction proved to be further challenging. The preferred method was comparison of power flow analysis solutions from the extracted and transformed data (in OpenDSS) to results in the original data format for the PSS/Sincal and PowerFactory network models. No such validation step was possible for the network data provided in excel, as it was not associated with any existing power flow analysis. Unfortunately, OpenDSS rarely gives very useful feedback if the data provided is inconsistent or incomplete. It tends to fail or run anyway and return incorrect results. Where extracted network models were validated, it was reasonably straightforward. However, there was a steep learning curve when it came to tracking down problems as it required experience with power flow, and a good understanding of the most likely origin of any mismatch – from incomplete, incorrect, or unusually expressed source data, through inadequate processing, to modelling approximations or inadequacies of the comparison power flow solver software.

3 The Process for clustering Low-Voltage Networks

This section describes the clustering process used to produce a selected set of low-voltage networks from the dataset provided, using the features derived as described in Chapter 2. Additional data curation was performed before a partition-based *k*-medoids clustering technique was applied across the feature set, eventually resulting in 23 clusters, each with an associated representative LV network.

We then show additional information about the types of LV networks across the dataset including the DNSPs whose LV networks belong to each cluster, key features characterising each cluster and the spread of clusters amongst regional characteristics. Finally, this section investigates some limitations with choices made in the data science process because of limitations with the data, and areas for improvement moving forward.

The LV network clustering process is presented in two flowcharts. Figure 3 shows the process of preparing the dataset for clustering and Figure 4 shows clustering to produce the final set of representative low-voltage networks. Section 3.1 explains the method of creating a cleaned dataset. Scaling is covered in Section 3.1.2 and dimensionality reduction is described in Section 3.1.3. Finally, clustering of the dimensionality reduced dataset is detailed in Section 3.2. Choosing the best performing set of clusters is covered in Section 3.3.

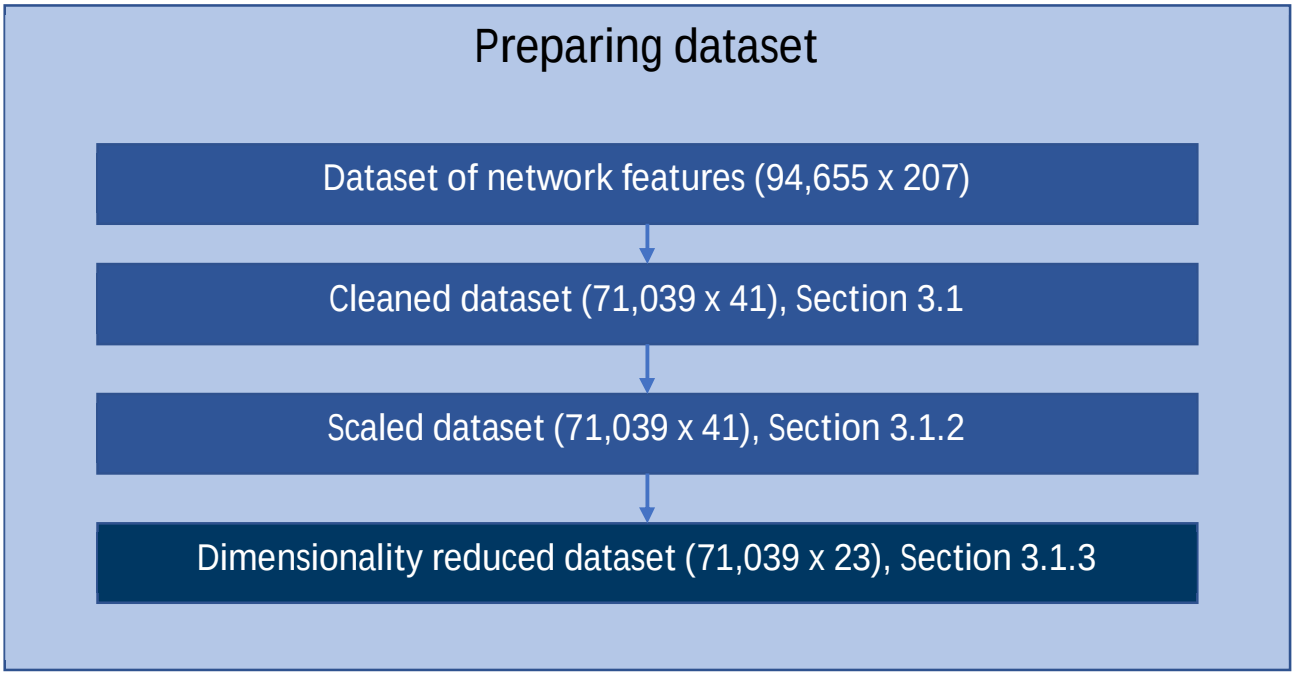


Figure 3 Diagram showing the preparation of the dataset for clustering

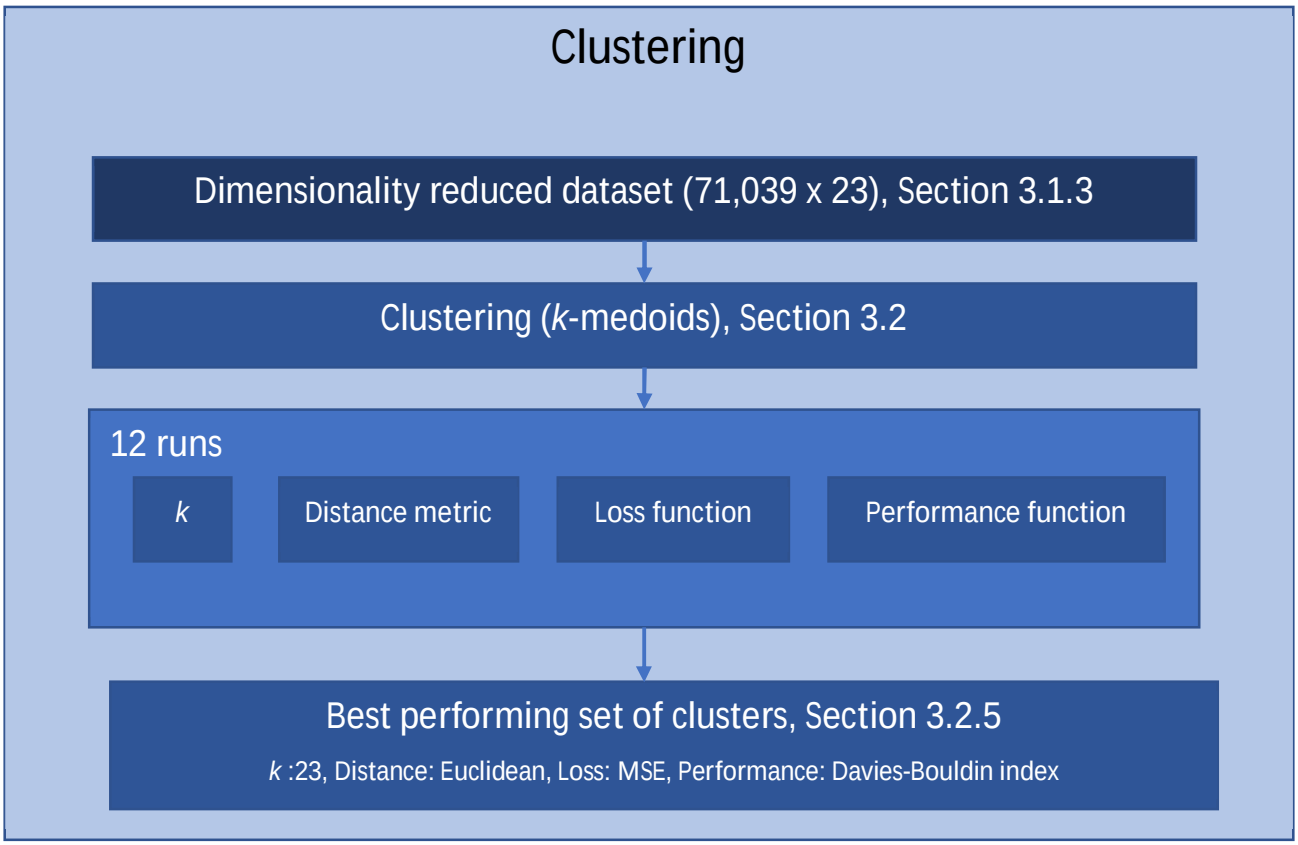


Figure 4 Diagram showing the clustering process beginning with the dimensionality reduced dataset

3.1 Data preparation

Before we run the clustering, the data is prepared in a suitable format**.** The network data was processed in PowerFactory and Sincal to produce the dataset in this section, which consists of 94,655 low-voltage networks and 207 features. Each *observation* is a low-voltage network, and each *feature* is a measurable value of that network (e.g. total line length, number of nodes, ratio of overhead lines, etc). The total list of features is found in Appendix A .

The data were sourced from seven DNSPs. There were two main issues in using this data:

1. The network counts are imbalanced across the DNSPs. Ausgrid and TasNetworks dominate the total number of networks while SAPN, Essential Energy, Energy Queensland and Ausnet are underrepresented. The number of networks is shown in Table 4.
2. Each DNSP records a distinct set of features for their networks. When a union is performed on the data across the DNSPs, missing data will be generated where one DNSP records a particular feature, whilst other DNSPs do not – and this accounts for most of the missing data which occurs. Since clustering relies upon complete data, any incomplete data needs to be selectively removed or interpolated.

To reduce the impact of missing data, we followed a manual process alternating between reducing networks and reducing features to reach a complete table of maximal size. Each feature was assessed independently to determine whether the feature needed to be omitted, or if there was a sufficiently small number of low-voltage networks with poor data quality, such that the networks can be dropped instead of the feature. First, the observations were dropped based on the feature

qualities. If it was later determined that the feature needed to be omitted then the observations were reconsidered for the final dataset.

Breaking down the final number of low-voltage networks and features:

- 23,616 *observations* (LV networks) were eliminated because they were an artefact of the source data:
    - 20 contained unlikely network cycles
    - 6,818 had a line length total of zero metres
    - 12,840 had an unreasonably low line length (<14m)
    - 36 had an unlikely high line length (>11,000m)
    - 7,498 low-voltage networks had no load information (3,902 of which had not already been removed already due to having an unrealistic line length).
- 166 *features* (columns) were deleted due to:
    - missing data (excess of 'NaN' values, 10 features dropped)
    - a standard deviation of zero (no data diversity, 10 dropped)
    - expert advice (combination of unimportant information and poor data quality, 139 dropped)
    - collinearity (1 column dropped)
    - being metadata (non-numerical data which was unable to be used for clustering, 6).

The cleaned dataset had 71,039 low-voltage networks and 41 features. A list of features is found in Appendix A . The split of networks across DNSP is seen in Table 4.

Table 4 Spread of low-voltage networks across Distribution Network Service Providers for both the raw and cleaned dataset.

	Raw dataset		Cleaned dataset	
DNSP	**Count**	**Proportion**	**Count**	**Proportion**
Ausgrid	60,182	0.63580	45,870	0.64570
TasNetworks	27,020	0.28546	23,357	0.32879
SAPN	4,723	0.04990	13	0.01246
Endeavour	1,633	0.01725	839	0.01181
Essential	1,017	0.01074	885	0.00058
Ausnet	46	0.00049	41	0.00048
EnergyQueensland	34	0.00036	34	0.00018
Total	**94,655**	**1**	**71,039**	**1**

3.1.1 Correlation analysis and feature selection

The cleaned dataset was checked for correlation to gauge how much information is contained within the features. A high correlation indicates redundant information, which could bias the clustering results. Clustering outcomes depend on the features selected for clustering, so reducing the number of features can lead to an improved set of clusters as well as an improvement upon the time and computational complexities.

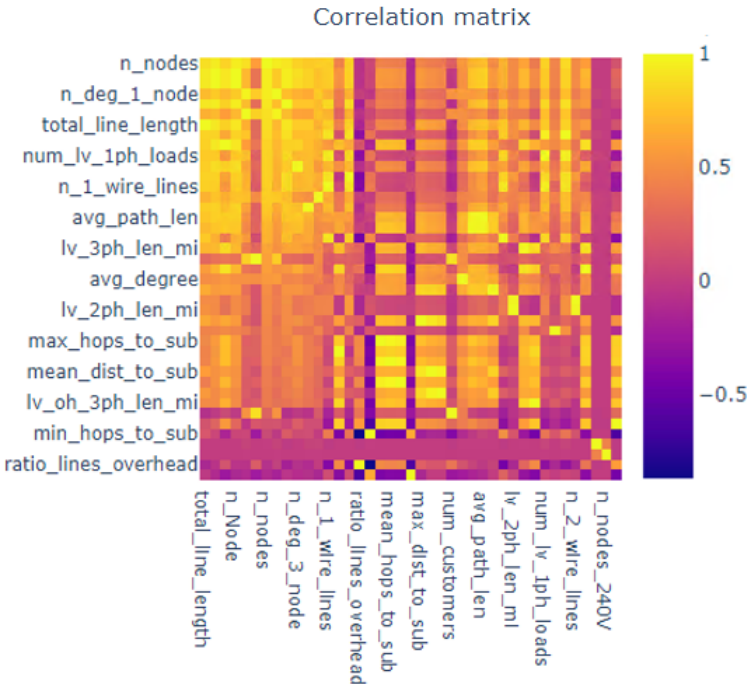


Figure 5 Correlation matrix showing the pairwise correlations of features in the dataset

The set of 41 features in the cleaned dataset have a high linear correlation. Figure 5 shows the correlation matrix of features in the dataset, while Table 5 shows the top 15 pairwise correlations. The high correlation shows that, although we have 41 features, the information the data contains could likely be summarised in fewer than 41 features. We opt to perform dimensionality reduction, however, before performing dimensionality reduction we need to scale the data.

Table 5 Top 15 pairwise correlations between features in the cleaned dataset (see Table 17 for explanation of feature descriptions).

Feature 1	Feature 2	Correlation coefficient
n_Line	n_Node	0.999
mean_hops_to_sub	median_hops_to_sub	0.990
max_hops_to_sub	mean_hops_to_sub	0.984
lv_2ph_len_mi	n_2_wire_lines	0.978
Diameter	avg_path_len	0.976
mean_dist_to_sub	median_dist_to_sub	0.975
n_Load	n_deg_1_node	0.967
num_lv_1ph_loads	avg_num_load_per_transformer	0.960
max_hops_to_sub	median_hops_to_sub	0.959
n_Line	n_deg_2_node	0.947
n_Node	n_deg_2_node	0.946
n_1_wire_lines	num_lv_1ph_loads	0.945
n_Node	n_nodes	0.946
n_Line	n_nodes	0.942

n_nodes	n_deg_1_node	0.939

3.1.2 Data scaling and normalisation

Scaling the data is an important data pre-processing step for dimensionality reduction and clustering. If we did not scale the data before processing, the clusters would be defined by the features with large magnitudes, and there would be little influence from the features with smaller magnitudes.

There are many ways to scale the data. In this project we scale each feature in the dataset to zero mean and unit variance (that is, convert each feature into z-scores). This ensures that the spread of the features is kept, and the clustering result does not depend on the units in which the features are expressed. The unit variance scaling also ensures that the within-feature relative spread and outliers are maintained without affecting the typical values as significantly as other type of scaling like min-max scaling[15].

3.1.3 Feature set dimensionality reduction

Dimensionality reduction is an optional step to prepare for clustering. Feature datasets are commonly sparse and describe the data with an excess of parameters. Running dimensionality reduction can reduce the sparsity and lead to an improvement in the quality of the clustering.

There are several techniques that can be used to reduce the number of dimensions. These are grouped into two categories: linear, and non-linear. Many of the features in this dataset relate to the topology of the network (e.g. number of nodes and wires, total line length, etc) which leads to a high linear correlation among features. Because of this we reduce the number of dimensions using Principal Component Analysis (PCA); the most common linear dimensionality reduction technique. PCA outperforms most non-linear dimensionality reduction techniques on real-life data (van der Maaten, Postma, & Herik, 2007).

PCA works by rearranging the 41 original features to create a new set of 41 features called principal components. The principal components are a linear combination of the original features, which are chosen by maximising variance in each successive principal component. The first principal components capture much of the information in the dataset, while later principal components typically hold minimal additional information. This is equivalent to capturing the latent variables in the dataset – features that cannot be measured directly – to reduce the sparsity of the dataset. The latter principal components can then be dropped with little information loss.

Figure 6 shows the proportion of information kept in the dataset by using PCA, with respect to the number of dimensions. Most of the information in the dataset is captured in a small number of dimensions. Over 99.5% of the information in the dataset is kept by dropping down to 23 dimensions. So, we reduce from 41 features to 23 dimensions using PCA.

[15] Min-max scaling is a method which scales a feature between [0, 1]. This scaling is highly sensitive to extreme values: the smallest value is mapped to 0 and the largest is mapped to 1. All other observations are mapped to their point on the scale.

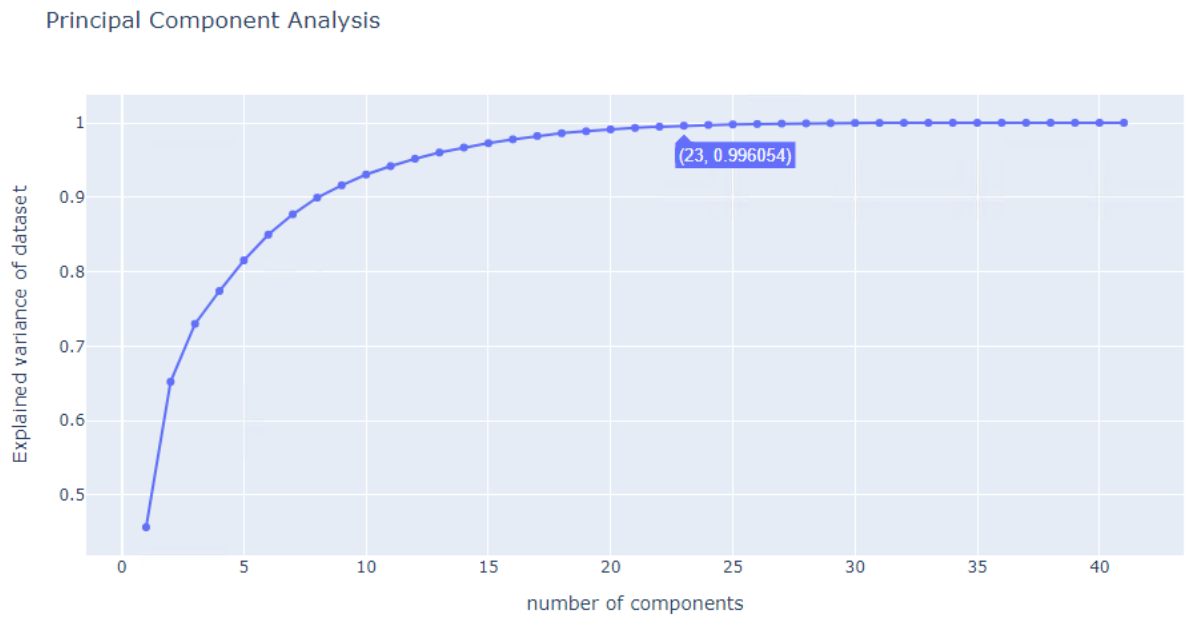


Figure 6 Variance of data explained versus number of components using Principal Component Analysis.

3.2 Clustering

With the dataset prepared for clustering, a representative set of low-voltage networks can be identified. There are many clustering techniques to choose, broadly grouped into five different categories. These are partition-based, density-based, hierarchical, grid-based, and model-based. The choice of clustering depends on the requirements of the problem being solved, in this case forming a set of representative networks which describe those actual networks in the dataset.

The solution that we chose for this project is partition-based *k*-medoids clustering. Many clustering methods use a set of mean feature values, however this would not be unlikely to correspond to any actual cluster member, contrary to the aim of this project. It is more important to produce actual networks representative of the clusters in each partition.

In *k*-medoid clustering an initial set of medoids is selected which are iteratively improved with the training data to form the final set of medoids. For the *k*-medoids partition-based clustering there are model hyperparameters to choose. These are:

- the number of clusters, k
- a distance metric (which measures the closeness between observations)
- a loss function (which gives an idea of how much error is in the system, and indicates how to improve the medoids in the next iteration)
- an initialisation state (which is the original choice of networks as medoids, before the training process - this affects the final choice of medoids)
- a performance-measuring function (which is used to help decide the final set of clusters).

Each parameter is explored in the following subsections. Adjusting the number of clusters k and the initial states create unique sets of medoids. From these medoids, the best performing set is selected by a performance-measuring function.

3.2.1 Choosing the number of clusters

The partition-based clustering requires us to specify a value of the number of clusters, k, to partition the space. This is one of the most important parameters for partition-based clustering and should be selected by a combination of data science techniques and expert analysis.

We trialled two different data science methods to find the value of k: the elbow-method and mean silhouette score. Both methods were run 100 times each yielding $k = 6$ clusters. Upon advice from domain experts, the value of k for the elbow method and mean silhouette score seemed too low to capture all the expected information about the low-voltage networks, so we trialled for $k = 2, 3, \dots, 40$ clusters, settling on $k = 23$.

3.2.2 Distance metric

Each clustering method requires a distance metric – a way to measure the closeness between two objects. The most common approach is the Euclidean metric which uses the Euclidean distance between observations. Because we performed PCA and represented the dataset by principal components, we used the Euclidean metric as it is invariant under translation and rotation. Many other distance metrics were not suitable as they are dependent on the raw set of features.

3.2.3 Loss function

The loss function is used for the training process. The loss function gives an idea of how much error is in the system and indicates how to improve the medoids in the next iteration. The choice of loss function will affect how the initial selection of medoids are tuned to represent each cluster. Because clustering is an unsupervised process, there is no fixed way to rule out the choice of any loss function. We trialled only Mean Squared Error (MSE) as a loss function because it is the most common and is easiest to compare with other studies done in this space.

3.2.4 Initialisation

The *k*-medoids algorithm is deterministic, but partition-based clustering depends on the initial random selection of medoids. Each time partition-based clustering randomly initializes a set of medoids and iteratively improves the set using the *k*-medoids algorithm. The final set of medoids will vary depending on the initial random state. This means that the clustering should be run multiple times to reduce the effect from chance. For each set of parameters, the clustering was run 12 times with different seeds to produce 12 sets of clusters which can be assessed to choose the best performing set.

3.2.5 Performance-measuring function

The performance-measuring function assesses the quality of the medoids by comparing the within-cluster variation to the amongst-cluster variation. We measure the performance by looking at four performance-measuring functions (see e.g. Halkidi, Batistakis, & Vazirgiannis (2001)):

- Mean Squared Error (MSE)
- Davies-Bouldin index (DB index)
- Calinski-Harabasz index (CH index)
- Mean silhouette coefficient.

A good score will show well-defined clusters which are separate to all other clusters while a poor score would have indistinguishable groups of clusters with high variance. The quality of the scores will be limited by the quality of the dataset. Clustering is an unsupervised technique, so the choice of performance-measuring function depends on a combination of data properties and expert opinion.

3.3 Choosing final network clusters

The choice of: number of clusters (k), performance-measuring function, and initialisation, can be summarised in Figure 7 which helps us to choose the final set of clusters. Each point is a single run of *k*-medoids clustering with PCA, with 23 components, Euclidean distance metric and MSE for the loss function. Of the four-performance metrics we chose the Davies-Bouldin Index. This is because previous unpublished clustering work in other domains has shown that low DB index scores tend to match well with our intuition around cluster quality, and because the DB index was able to identify a clear 'elbow point' (sharp improvement in score compared to surrounding points) at $k = 23$. Additionally, domain experts examined the cluster medoids and associated data from the best runs (red points in Figure 7) and were satisfied that the final $k = 23$ result did not show any obvious peculiarities.

Clustering runs with lower values ($k < 16$) were dismissed as they were observed to select clusters that appeared primarily to merely split the dataset into clusters consisting almost exclusively of networks from one of two DNSPs, partly due to the highly imbalanced dataset. Clustering with higher-k values were more likely to create clusters with additional network configurations from the smaller DNSP sets that were considered interesting for subsequent power flow simulations. A side-effect of the higher-k results was that a few clusters ended up containing only very small numbers of low-voltage networks. This was considered a reasonable trade-off for the additional clusters that appeared as a consequence, with medoid LV networks that were more interesting.

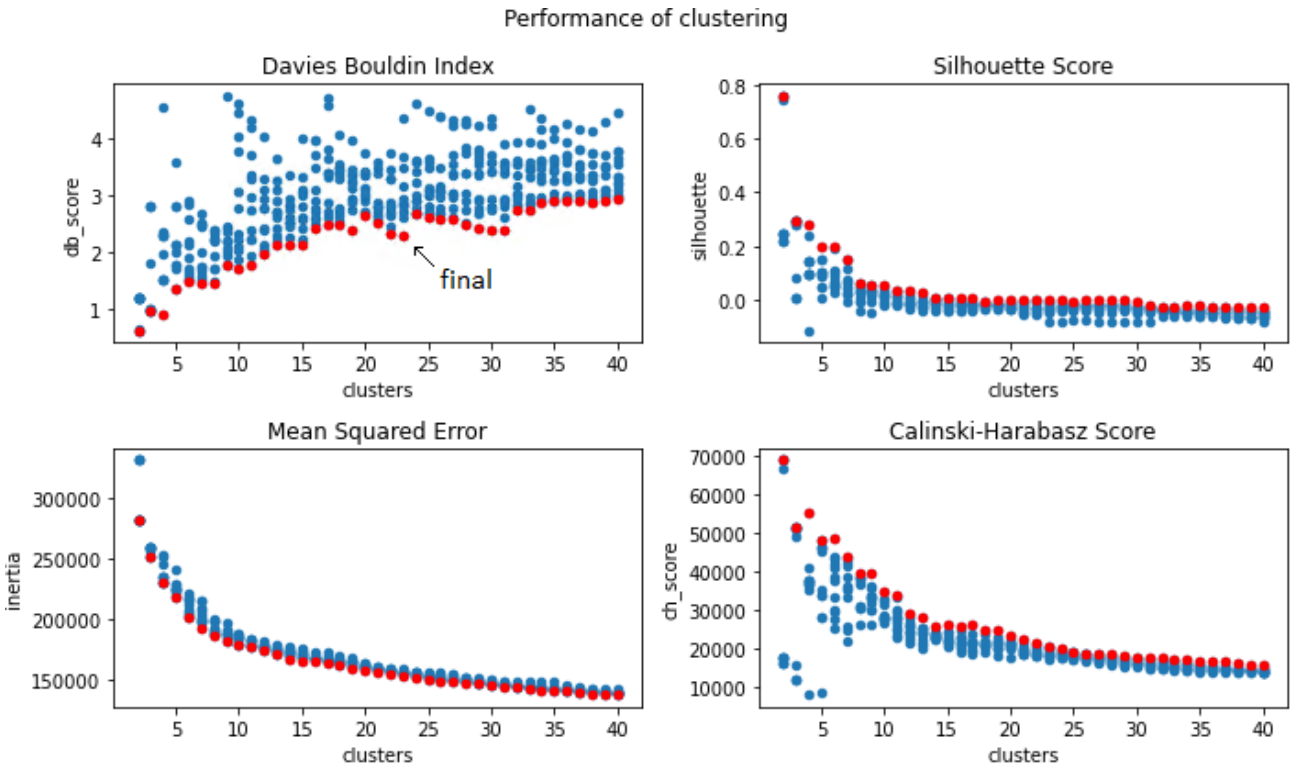


Figure 7 Results from clustering with cluster number k = 2, 3, ..., 40 for four performance-measuring functions, with 12 runs per choice of *k*

The result of each run depends on the random initialisation of medoids, so multiple runs were performed for each k to minimise the influence of the initial random seed. Each vertical stack of points shows cluster quality metrics for a single value of k with different random initialisation seeds. Red points show the best run for each stack (higher is better for the silhouette score and CH Index, and lower is better for the DB Index and MSE). Blue points show other runs.

Using Figure 7 we decide on the final set of representative networks. We use the Davies-Bouldin index and find an improvement at k=22 and k=23. The latter has runs with an improved score and is chosen as the final set of networks which will be used in the rest of this report.

3.4 Limitations of clusters

There were two main limitations with this dataset which biased the final set of clusters.

1. The starting dataset had missing data. When we dropped the *observations* and *features*, we best reduced the bias to ensure that maximal information is kept. This comes at the cost of losing load and impedance information.
2. There are limitations with the original dataset due to a bias towards south-eastern Australian low-voltage networks. We treated all networks as independent in the clustering, so this bias exists with the final set of clusters.

A final limitation is to be noted. The clusters identified, reported on, and from which cluster representative medoids are selected are based on incorrectly parsed feeder data, which missed some loads. When corrected, the new medoids appeared similar to those identified using the erroneous parsing process. Consequently, the originally identified ones were retained. Details appear in Appendix D .

4 Clustering Results Discussion

This section investigates the results of the clustering process in more depth and explores what it tells us about typical LV networks. The selected clustering run produced 23 clusters. To visualise the performance across two of the raw features in the dataset, Figure 8 shows a scatter plot of total line length vs the number of nodes, colour-coded by the cluster. This shows that these two features explain the identified clusters to a large extent.

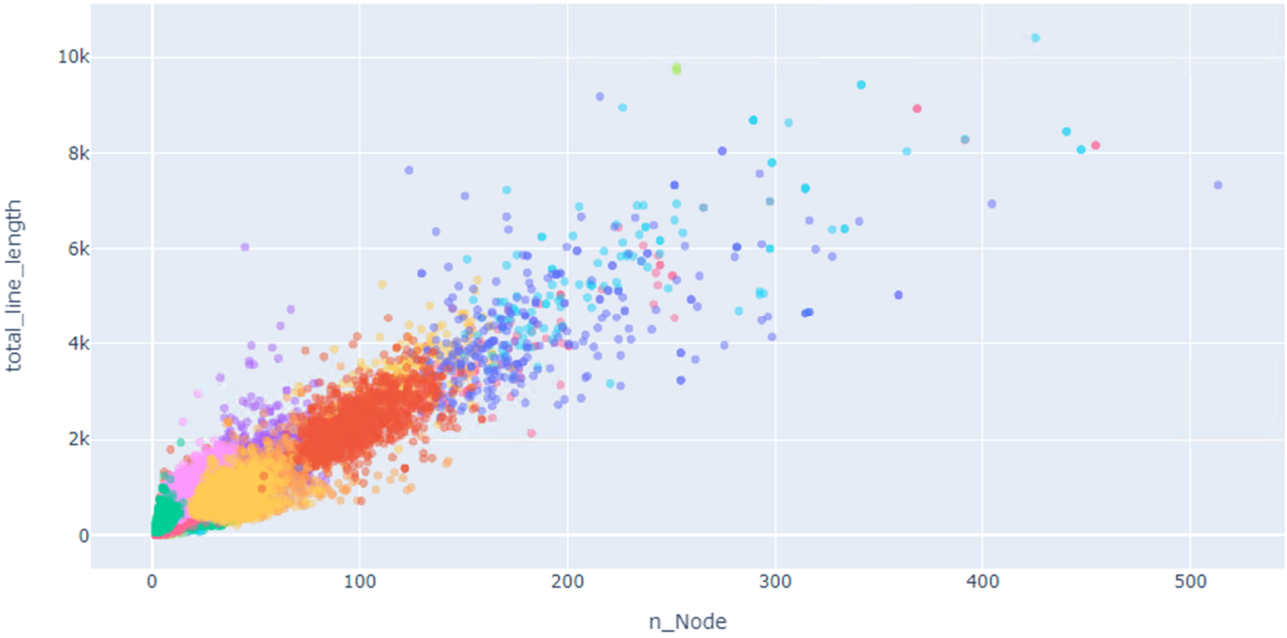


Figure 8 A scatter plot of two important clustering features: [total_line_length] vs [n_Node]. Each colour indicates a distinct cluster

4.1 Feature importance

We used the cluster labels from the final set of clusters to train a supervised binary classifier (random forest model), from which the estimated feature importance could be extracted. For each cluster, the network cluster label was set to one if the network belongs to the cluster and zero if the network does not belong to the cluster. Each classifier was trained on 80% of the data with the remaining 20% used for verifying that the classifiers' prediction accuracy was high enough to trust the results. Essentially, this estimates the importance of each raw clustering feature for each cluster with respect to all other clusters, giving useful insight into what caused each cluster to be selected.

4.2 Cluster analysis

The following pages show an analytical breakdown of the clusters and their key features.

Table 6 shows the renders of the 23 final cluster medoids. The remoteness area information determines the row and the size of the low-voltage network (in terms of total line length) is ascending along each row from left to right. The colour of each cell in the table indicates whether the mesh block category type is residential or primary production, with location (mesh block category) on the vertical axis, and size (total line length in metres) on the horizontal axis. Networks

situated primarily in residential areas are shown with blue backgrounds, and primary production (i.e. farming areas) with green backgrounds. To help visually interpret the network renders, lines and nodes are coloured such that:

- power sources are always red
- feeder-head nodes are always yellow
- thicker lines are feeders (heuristically-identified, see Section 2.1.3)
- thin lines are non-feeders.

All other colours are arbitrarily (based on their order-of-discovery) assigned from a fixed colour-map based on:

- the node class (for nodes)
- distinct combination of properties for each line – by default: impedance per unit length, cable type, line category (as classified by the DNSP), and nominal voltage, or a subset if not all properties are available.

Line colours are usually consistent across a single DNSP's networks because the ordering of the lines in the source data is preserved, but this is coincidental rather than guaranteed.

Table 6 Renders of the 23 final cluster medoids

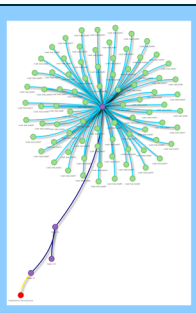
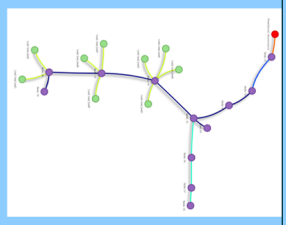
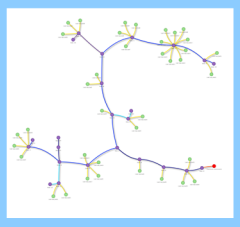
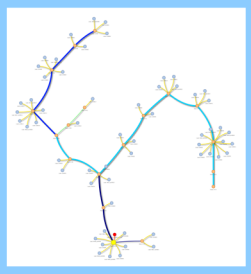
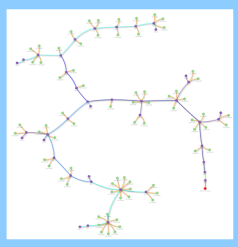
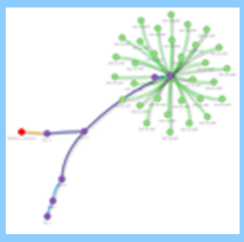
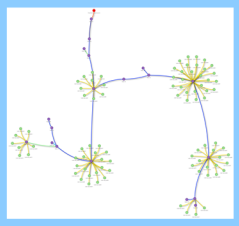
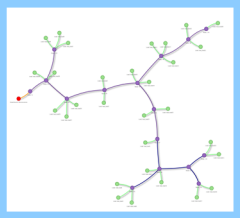
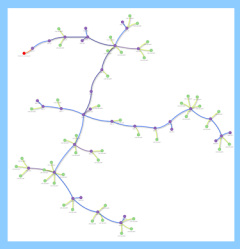
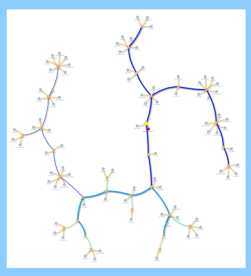
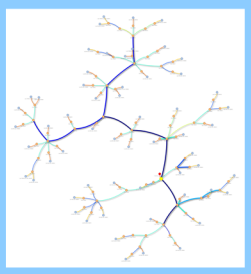
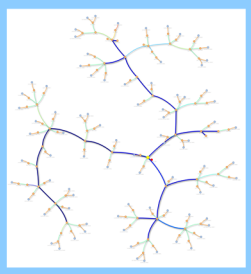
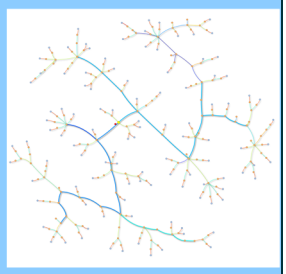
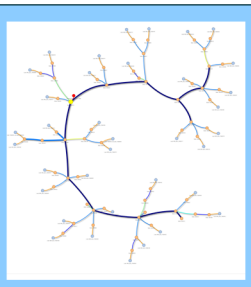
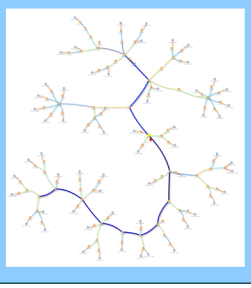
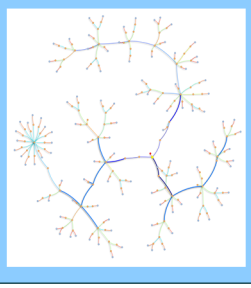
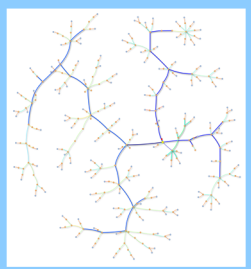
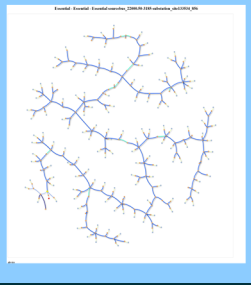
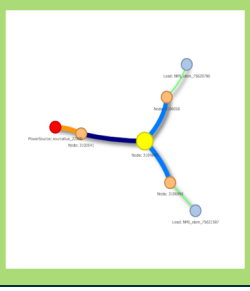
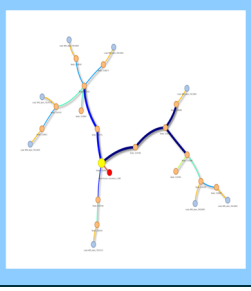
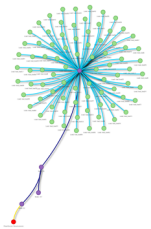
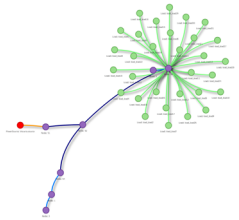
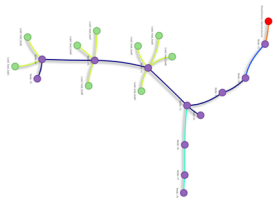
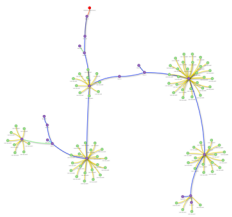
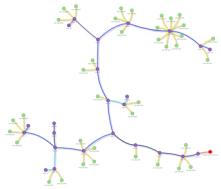
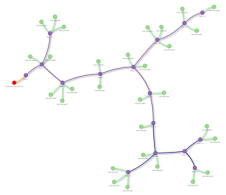
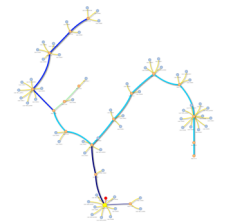
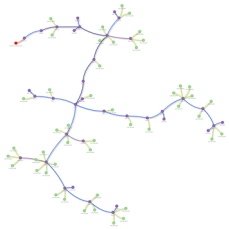
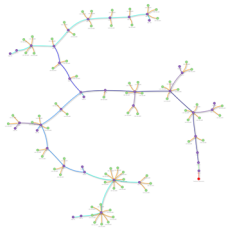
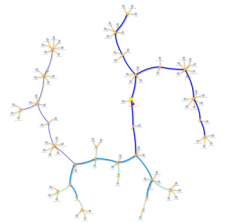
	Increasing network size (total line length) →				
City					
					
Suburban					
					
Regional					
					

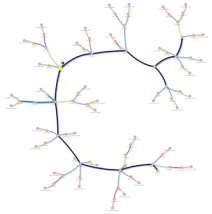
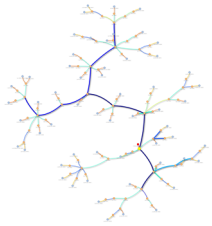
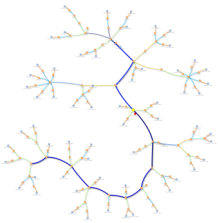
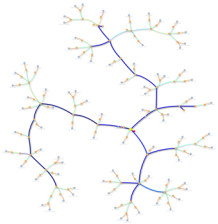
Table 7 shows individual network information about each of the clusters including the top features (and their values) which influenced the clustering as described in Section 3.2. It contains information about the number of networks that belong to the cluster, as well as a breakdown by DNSP. This table shows the top five most important features average values and importance estimates (higher value means more important) per cluster, the geographical categories (Meshblock and Remoteness Area) of the cluster's medoid, and how many networks fell into each cluster in total and per DNSP. Table 8 shows the full set of feature importance for each cluster. Table 10 shows the estimated relative importance of all the features used to obtain the final clustering result, averaged over all the clusters.

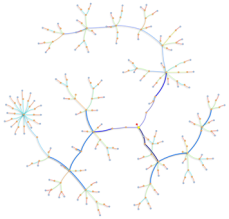
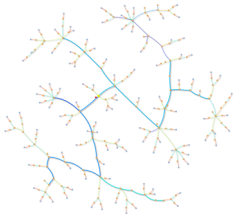
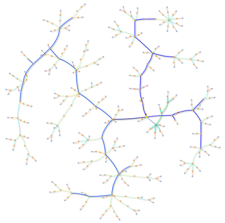
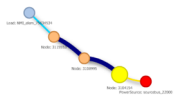
Table 7 Cluster snapshots

Label	Medoid Rendering	Most Important Features & Average Values	Meshblock Category	Remoteness Area Category	Network count	Networks in cluster per DNSP
A		feature / Value / importance num_customers 89.8 0.86 mean_hops_to_sub 3 0.026 n_overhead_lines 5.7 0.015 n_Load 90.2 0.013 n_nodes_400V 6.7 0.013	Residential	City	562	Ausgrid 561 Ausnet 0 Endeavour 1 EnergyQueensland 0 Essential 0 SAPN 0 TasNetworks 0
B		feature / value / importance num_customers 30.2 0.251 mean_hops_to_sub 3.1 0.148 lv_3ph_len_mi 0.1 0.128 lv_oh_3ph_len_mi 0.1 0.084 total_line_length 126.8 0.06	Residential	City	2197	Ausgrid 2179 Ausnet 0 Endeavour 18 EnergyQueensland 0 Essential 0 SAPN 0 TasNetworks 0
C		feature / value / importance n_nodes 17.8 0.318 mean_dist_to_sub 110 0.224 n_nodes_400V 11.7 0.128 mean_hops_to_sub 4.9 0.066 avg_path_len 3.1 0.049	Residential	City	10314	Ausgrid 10243 Ausnet 1 Endeavour 64 EnergyQueensland 2 Essential 0 SAPN 0 TasNetworks 4

Label	Medoid Rendering	Most Important Features & Average Values	Meshblock Category	Remoteness Area Category	Network count	Networks in cluster per DNSP
D		**feature** / **value** / **importance** num_customers 30 0.427 mean_hops_to_sub 6.5 0.116 total_line_length 372.7 0.086 lv_oh_3ph_len_mi 0.2 0.049 lv_3ph_len_mi 0.2 0.045	**Residential**	**City**	**3433**	**Ausgrid** 3430 **Ausnet** 0 **Endeavour** 3 **EnergyQueensland** 0 **Essential** 0 **SAPN** 0 **TasNetworks** 0
E		**feature** / **value** / **importance** n_nodes_400V 24.2 0.307 mean_dist_to_sub 147.7 0.151 n_nodes 33.6 0.114 mean_hops_to_sub 7.5 0.086 avg_path_len 4.7 0.039	**Residential**	**City**	**8435**	**Ausgrid** 8430 **Ausnet** 0 **Endeavour** 3 **EnergyQueensland** 0 **Essential** 0 **SAPN** 0 **TasNetworks** 2
F		**feature** / **value** / **importance** n_nodes 21.8 0.283 mean_dist_to_sub 215.7 0.166 n_nodes_400V 16.4 0.126 lv_oh_3ph_len_mi 0.3 0.077 mean_hops_to_sub 6 0.063	**Residential**	**City**	**4505**	**Ausgrid** 4482 **Ausnet** 1 **Endeavour** 7 **EnergyQueensland** 0 **Essential** 3 **SAPN** 0 **TasNetworks** 12
G		**feature** / **value** / **importance** n_lines_240V 20.7 0.873 diameter 14.4 0.046 n_PhaseLoad 61.3 0.025 n_deg_>4_node 6.3 0.014 n_Load 61.1 0.013	**Residential**	**City**	**10**	**Ausgrid** 0 **Ausnet** 0 **Endeavour** 0 **EnergyQueensland** 0 **Essential** 0 **SAPN** 10 **TasNetworks** 0

Label	Medoid Rendering	Most Important Features & Average Values	Meshblock Category	Remoteness Area Category	Network count	Networks in cluster per DNSP
H		**feature** / **value** / **importance** n_nodes_400V — 35.5 — 0.487 mean_hops_to_sub — 9.1 — 0.089 max_dist_to_sub — 377.8 — 0.089 max_sub_node_distance_mi — 0.2 — 0.059 mean_dist_to_sub — 200.9 — 0.033	**Residential**	**City**	**6055**	**Ausgrid** 6045 **Ausnet** 2 **Endeavour** 1 **EnergyQueensland** 4 **Essential** 2 **SAPN** 0 **TasNetworks** 1
I		**feature** / **value** / **importance** mean_hops_to_sub — 11.1 — 0.383 lv_oh_3ph_len_mi — 0.6 — 0.161 max_sub_node_distance_mi — 0.3 — 0.097 mean_dist_to_sub — 277 — 0.067 n_nodes_400V — 45.4 — 0.05	**Residential**	**City**	**1594**	**Ausgrid** 1581 **Ausnet** 0 **Endeavour** 2 **EnergyQueensland** 0 **Essential** 9 **SAPN** 0 **TasNetworks** 2
J		**feature** / **value** / **importance** n_nodes_240V — 3 — 0.818 n_lines_240V — 30 — 0.182	**Residential**	**City**	**1**	**Ausgrid** 0 **Ausnet** 0 **Endeavour** 0 **EnergyQueensland** 0 **Essential** 0 **SAPN** 1 **TasNetworks** 0
K		**feature** / **value** / **importance** max_sub_node_distance_mi — 0 — 0.237 n_nodes — 8 — 0.158 median_dist_to_sub — 31.1 — 0.142 avg_path_len — 2 — 0.111 mean_dist_to_sub — 35.6 — 0.09	**Primary production**	**Suburban**	**9394**	**Ausgrid** 8919 **Ausnet** 1 **Endeavour** 298 **EnergyQueensland** 0 **Essential** 157 **SAPN** 0 **TasNetworks** 19

Label	Medoid Rendering	Most Important Features & Average Values	Meshblock Category	Remoteness Area Category	Network count	Networks in cluster per DNSP
L		feature / value / importance avg_num_load_per_transformer 34 0.363 n_nodes 90.4 0.326 n_Node 53.9 0.101 n_PhaseLoad 39 0.041 n_Line 53.6 0.037	Residential	Suburban	1771	Ausgrid 0 Ausnet 2 Endeavour 104 EnergyQueensland 2 Essential 0 SAPN 0 TasNetworks 1663
M		feature / value / importance num_lv_3ph_loads 34.5 0.577 n_Wire 202.8 0.193 n_nodes 148.7 0.057 n_PhaseLoad 132 0.051 avg_degree 2 0.019	Residential	Suburban	342	Ausgrid 0 Ausnet 0 Endeavour 23 EnergyQueensland 0 Essential 3 SAPN 0 TasNetworks 316
N		feature / value / importance n_1_wire_lines 69.5 0.267 n_nodes 171.7 0.156 num_lv_1ph_loads 66.4 0.145 n_Node 100.4 0.098 lv_3ph_len_mi 0.6 0.06	Residential	Suburban	946	Ausgrid 0 Ausnet 5 Endeavour 136 EnergyQueensland 0 Essential 0 SAPN 0 TasNetworks 805
O		feature / value / importance n_2_wire_lines 37.4 0.492 n_nodes 191.8 0.216 lv_2ph_len_mi 0.8 0.098 n_Node 111.9 0.065 n_underground_lines 105.1 0.032	Residential	Suburban	335	Ausgrid 0 Ausnet 0 Endeavour 0 EnergyQueensland 0 Essential 0 SAPN 0 TasNetworks 335

Label	Medoid Rendering	Most Important Features & Average Values	Meshblock Category	Remoteness Area Category	Network count	Networks in cluster per DNSP
P		**feature** / **value** / **importance** num_lv_3ph_loads — 62.9 — 0.506 n_Wire — 394.3 — 0.195 n_3_wire_lines — 104.3 — 0.064 n_nodes — 285.6 — 0.03 n_Node — 166.4 — 0.024	**Residential**	**Suburban**	**133**	**Ausgrid** 0 **Ausnet** 0 **Endeavour** 4 **EnergyQueensland** 0 **Essential** 2 **SAPN** 0 **TasNetworks** 127
Q		**feature** / **value** / **importance** n_1_wire_lines — 120.2 — 0.375 lv_2ph_len_mi — 0.1 — 0.146 n_nodes — 311.4 — 0.118 n_Node — 188.3 — 0.055 n_3_wire_lines — 62.7 — 0.049	**Residential**	**Suburban**	**288**	**Ausgrid** 0 **Ausnet** 27 **Endeavour** 38 **EnergyQueensland** 0 **Essential** 0 **SAPN** 0 **TasNetworks** 223
R		**feature** / **value** / **importance** lv_2ph_len_mi — 1.5 — 0.448 n_2_wire_lines — 77.7 — 0.178 n_nodes — 377.6 — 0.143 n_deg_1_node — 170.8 — 0.114 n_1_wire_lines — 131 — 0.017	**Residential**	**Suburban**	**123**	**Ausgrid** 0 **Ausnet** 0 **Endeavour** 0 **EnergyQueensland** 0 **Essential** 0 **SAPN** 0 **TasNetworks** 123
S		**feature** / **value** / **importance** n_PhaseLoad — 1.9 — 0.447 mean_dist_to_sub — 47.9 — 0.29 min_dist_to_sub — 31.8 — 0.067 total_line_length — 86.4 — 0.038 ratio_lines_overhead — 0.1 — 0.035	**Primary Production**	**Regional**	**8944**	**Ausgrid** 0 **Ausnet** 1 **Endeavour** 0 **EnergyQueensland** 12 **Essential** 376 **SAPN** 0

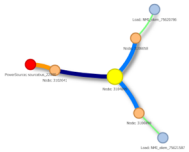
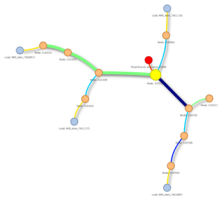
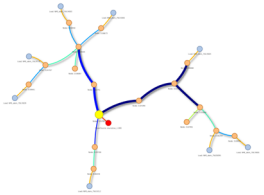
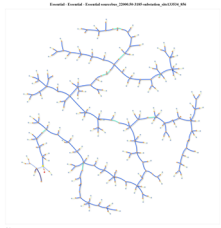
Label	Medoid Rendering	Most Important Features & Average Values	Meshblock Category	Remoteness Area Category	Network count	Networks in cluster per DNSP
T		feature / value / importance min_dist_to_sub 98.9 0.61 avg_path_len 2.5 0.13 mean_dist_to_sub 132.4 0.079 total_line_length 231.6 0.028	Primary Production	Regional	2770	Ausgrid 0 Ausnet 0 Endeavour 0 EnergyQueensland 4 Essential 171 SAPN 0 TasNetworks 2595
U		feature / value / importance total_line_length 376 0.314 n_underground_lines 7 0.216 avg_path_len 3.5 0.109 min_dist_to_sub 30.2 0.093 mean_dist_to_sub 106.8 0.07	Primary Production	Regional	5233	Ausgrid 0 Ausnet 1 Endeavour 74 EnergyQueensland 2 Essential 116 SAPN 1 TasNetworks 5039
V		feature / value / importance n_nodes 30.5 0.214 total_line_length 839.1 0.165 n_1_wire_lines 13.8 0.157 lv_1ph_len_mi 0.4 0.113 mean_hops_to_sub 4.1 0.078	Residential	Regional	3648	Ausgrid 0 Ausnet 0 Endeavour 63 EnergyQueensland 8 Essential 40 SAPN 1 TasNetworks 3536
W		feature / value / importance lv_oh_3ph_len_mi 5.9 0.421 lv_3ph_len_mi 6.1 0.316 total_line_length 9749.1 0.053 n_overhead_lines 249 0.053 mean_dist_to_sub 805.2 0.053 max_sub_node_distance_mi 1 0.053 num_lv_3ph_loads 136 0.053	Residential	Regional	6	Ausgrid 0 Ausnet 0 Endeavour 0 EnergyQueensland 0 Essential 6 SAPN 0 TasNetworks 0

Table 8 Relative importance for clustering (higher values means more important) of each network feature, by cluster; shading is scaled to each column (cluster)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
total_line_length	0.003	0.06	0.022	0.086	0.023	0.013	0.006	0.01	0.013	0	0.014	0.007	0.008	0.019	0.003	0.014	0.014	0.004	0.038	0.028	0.314	0.165	0.053
n_Line	0	0.002	0.003	0.031	0.012	0.005	0	0.012	0.002	0	0	0.037	0.002	0.009	0.013	0.011	0.003	0.013	0.001	0.002	0.003	0.006	0
n_Wire	0.008	0.042	0.003	0.006	0.011	0.008	0	0.013	0.003	0	0	0.005	0.193	0.018	0.005	0.195	0.003	0.004	0.001	0.001	0.003	0.008	0
n_Node	0.004	0.03	0.002	0.012	0.009	0.005	0	0.012	0.002	0	0	0.101	0.002	0.098	0.065	0.024	0.055	0.014	0.001	0.01	0.004	0.005	0
n_Load	0.013	0.054	0.004	0.023	0.019	0.002	0.013	0.005	0.005	0	0.001	0.001	0.001	0.001	0.001	0.002	0.004	0.001	0.001	0.001	0.002	0.001	0
n_PhaseLoad	0.009	0.048	0.004	0.023	0.013	0.003	0.025	0.004	0.002	0	0.002	0.041	0.051	0.019	0.004	0.006	0.013	0.001	0.447	0.002	0.001	0.002	0
n_nodes	0.001	0.004	0.318	0.006	0.114	0.283	0	0.023	0.02	0	0.158	0.326	0.057	0.156	0.216	0.03	0.118	0.143	0.003	0.009	0.006	0.214	0
n_deg_1_node	0.002	0.002	0.005	0.012	0.012	0.004	0	0.009	0.005	0	0.003	0.003	0.003	0.003	0.015	0.01	0.002	0.114	0	0.002	0.002	0.004	0
n_deg_2_node	0	0.002	0.002	0.005	0.006	0.006	0.01	0.006	0.008	0	0.001	0.001	0.003	0.003	0.002	0	0.003	0.011	0.001	0.003	0.003	0.002	0
n_deg_3_node	0.003	0.002	0.002	0.004	0.004	0.003	0	0.003	0.006	0	0	0.003	0.001	0.002	0.002	0.011	0.009	0.001	0	0.001	0.002	0.003	0
n_deg_4_node	0	0.001	0.001	0.002	0.004	0.002	0	0.003	0.002	0	0	0.003	0.002	0.004	0.003	0	0.006	0	0	0.001	0.001	0.002	0
n_deg_>4_node	0.001	0.001	0.001	0.005	0.004	0.002	0.014	0.002	0.001	0	0.001	0.001	0.004	0.001	0.002	0.01	0.003	0	0	0	0	0.001	0
n_1_wire_lines	0	0	0	0	0	0	0	0	0	0	0	0.002	0.012	0.267	0.002	0.002	0.375	0.017	0.001	0.004	0.007	0.157	0
n_overhead_lines	0.015	0.006	0.003	0.016	0.013	0.005	0	0.011	0.005	0	0.03	0.002	0.002	0.003	0.001	0	0.004	0	0.026	0.008	0.005	0.008	0.053
n_underground_lines	0	0	0	0	0	0	0.008	0	0.001	0	0.046	0.008	0.001	0.003	0.032	0.009	0.002	0.001	0.001	0.007	0.216	0.005	0
ratio_lines_overhead	0	0	0	0	0	0	0	0	0	0	0.03	0.001	0.001	0.002	0	0.001	0.001	0.001	0.035	0.02	0.004	0.004	0
min_hops_to_sub	0	0	0	0	0	0	0	0	0.004	0	0.003	0	0	0	0	0	0	0	0.005	0	0	0	0
max_hops_to_sub	0.001	0.005	0.002	0.005	0.007	0.006	0	0.004	0.019	0	0.001	0.001	0.005	0.002	0.001	0.008	0.001	0.001	0.01	0.001	0.005	0.003	0
mean_hops_to_sub	0.026	0.148	0.066	0.116	0.086	0.063	0	0.089	0.383	0	0.006	0.003	0.003	0.002	0.001	0.007	0.003	0.002	0.001	0.004	0.025	0.078	0
median_hops_to_sub	0.005	0.005	0.004	0.003	0.005	0.007	0	0.009	0.008	0	0.002	0.001	0.001	0	0.001	0.004	0	0	0	0.001	0.001	0.001	0
min_dist_to_sub	0	0	0	0	0	0	0.005	0	0	0	0.066	0.003	0.002	0.002	0.004	0.002	0.004	0.001	0.067	0.61	0.093	0.009	0
max_dist_to_sub	0.005	0.017	0.018	0.011	0.025	0.019	0	0.089	0.02	0	0.008	0.002	0.002	0.004	0.002	0.006	0.002	0	0.007	0.019	0.012	0.018	0
mean_dist_to_sub	0.004	0.029	0.224	0.012	0.151	0.166	0	0.033	0.067	0	0.09	0.002	0.002	0.004	0.001	0.001	0.002	0.001	0.29	0.079	0.07	0.065	0.053
median_dist_to_sub	0.002	0.01	0.012	0.008	0.017	0.049	0	0.022	0.021	0	0.142	0.003	0.004	0.003	0.001	0.001	0.003	0	0.007	0.02	0.014	0.01	0
num_customers	0.86	0.251	0.007	0.427	0.015	0.004	0	0.004	0.013	0	0.001	0	0	0	0	0	0	0	0	0	0	0	0
avg_degree	0.001	0.003	0.009	0.004	0.022	0.03	0	0.007	0.018	0	0.02	0.006	0.019	0.019	0	0.004	0.006	0.004	0.011	0.009	0.009	0.005	0
diameter	0.003	0.006	0.013	0.013	0.011	0.015	0.046	0.012	0.016	0	0.002	0.001	0.001	0.003	0.002	0.005	0.001	0.001	0.001	0.001	0.007	0.008	0
avg_path_len	0.006	0.014	0.049	0.022	0.039	0.053	0	0.027	0.021	0	0.111	0.005	0	0.007	0.003	0.014	0.008	0.002	0.03	0.13	0.109	0.022	0
max_sub_node_distance_mi	0.002	0.019	0.048	0.02	0.019	0.022	0	0.059	0.097	0	0.237	0.003	0.001	0.004	0.002	0.001	0.01	0	0.006	0.016	0.027	0.045	0.053
lv_1ph_len_mi	0	0	0	0	0	0	0	0	0.001	0	0	0.004	0.001	0.006	0.006	0.003	0.028	0.016	0.005	0.006	0.033	0.113	0
lv_2ph_len_mi	0	0	0	0	0	0	0	0	0	0	0.003	0.009	0.003	0.017	0.098	0.016	0.146	0.448	0	0.001	0.007	0.003	0
lv_3ph_len_mi	0.004	0.128	0.017	0.045	0.021	0.014	0	0.01	0.009	0	0.01	0.004	0.015	0.06	0.006	0.007	0.024	0.005	0.001	0.001	0.002	0.004	0.316
lv_oh_3ph_len_mi	0.002	0.084	0.026	0.049	0.018	0.077	0	0.025	0.161	0	0.008	0.006	0.002	0.004	0.002	0.002	0.004	0.003	0	0.001	0.003	0.005	0.421
num_lv_1ph_loads	0.001	0	0	0	0	0	0	0	0.002	0	0	0.022	0.002	0.145	0.005	0.003	0.013	0.008	0.001	0.002	0.002	0.002	0
num_lv_3ph_loads	0	0	0	0	0	0	0	0	0	0	0	0.011	0.577	0.025	0	0.506	0.039	0	0	0	0.001	0.001	0.053
avg_num_load_per_transformer	0	0	0	0	0	0.001	0	0	0.007	0	0	0.363	0.003	0.001	0.004	0.004	0.004	0.002	0.001	0.001	0.004	0.018	0
n_2_wire_lines	0	0	0	0	0	0	0	0	0	0	0	0.005	0.003	0.042	0.492	0.016	0.036	0.178	0.001	0	0	0.001	0
n_3_wire_lines	0.007	0.013	0.002	0.02	0.012	0.007	0	0.011	0.003	0	0	0.003	0.012	0.042	0.006	0.064	0.049	0.001	0	0	0.002	0.002	0
n_lines_240V	0	0	0	0	0	0	0.873	0	0	0.182	0	0	0	0	0	0	0	0	0	0	0	0	0
n_nodes_240V	0	0	0	0	0	0	0	0	0	0.818	0	0	0	0	0	0	0	0	0	0	0	0	0
n_nodes_400V	0.013	0.012	0.128	0.014	0.307	0.126	0	0.487	0.05	0	0.001	0	0	0	0	0	0	0	0	0	0	0	0

Table 9 Mean values for features, by cluster; shading is scaled to each row

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
total_l ine_length	132.7	126.8	288	372.6	489	564	857.5	756.7	1040.1	1126	85.59	1454	1942.3	2376.6	2973.2	3776.4	4323.8	5644.3	86.42	231.5	375.9	839.1	9749
n_Line	5.665	5.383	10.78	18.43	23.19	15.48	20.7	34.57	44.753	30	3.984	53.597	85.269	100.79	112.26	170	190.51	235.52	2.493	3.227	8.875	19.07	252
n_Wire	17.16	16.16	32.34	55.56	69.75	46.45	62.1	103.9	134.87	90	11.86	87.093	202.76	159.77	162.63	394.25	325.87	362.22	3.336	4.075	12.75	28.73	762
n_Node	6.665	6.383	11.78	19.43	24.19	16.47	21.7	35.54	45.75	31	4.984	53.851	84.436	100.37	111.88	166.36	188.29	231.11	3.471	4.2	9.751	19.82	253
n_Load	90.19	30.62	5.018	30.02	8.417	4.329	61.1	9.013	10.168	87	2	34.627	62.287	69.489	77.958	117.21	121.2	144.49	1.625	1.881	4.389	8.728	136
n_PhaseLoad	359.6	121.2	19.86	120	33.64	17.24	61.3	35.95	40.327	87	7.767	39.018	132.02	75.373	87.809	244.78	132.27	183.05	1.95	2.193	4.98	9.892	408
n_nodes	97.85	38	17.8	50.45	33.61	21.8	84.8	45.56	56.925	120	8.004	90.418	148.65	171.71	191.84	285.55	311.36	377.61	7.095	8.079	16.12	30.52	391
n_deg_1_node	92.73	33.16	8.992	36.68	16.25	9.734	63	20.85	25.101	88	4.034	38.364	66.494	74.45	83.101	128.31	136.29	170.78	2.959	3.166	6.263	11.31	149.667
n_deg_2_node	2.916	2.688	5.64	7.792	10.81	7.773	2.9	15.18	19.435	5	2.862	33.339	52.365	62.297	70.89	103.13	108.53	135.21	3.389	3.997	7.14	13.01	149.333
n_deg_3_node	0.847	0.726	1.772	3.139	3.787	2.769	3.7	5.865	8.354	4	0.688	10.271	15.091	17.763	17.519	25.692	35.17	33.447	0.602	0.749	1.828	4.183	50.5
n_deg_4_node	0.148	0.129	0.639	0.768	1.282	0.797	5.2	1.953	2.24	9	0.227	4.062	5.488	7.239	9.134	9.955	14.729	14.537	0.103	0.118	0.577	1.317	29.333
n_deg_>4_node	1.176	1.252	0.476	1.83	0.96	0.401	6.3	1.026	1.078	9	0.106	1.993	5.298	4.841	5.233	11.128	8.396	13.187	0.013	0.018	0.13	0.214	2
n_1_wire_lines	0.002	0.017	0.034	0.003	0.005	0.026	0	0.007	0.074	0	0.044	35.471	25.766	69.517	67.8	55.173	120.222	131.033	2.017	2.704	6.774	13.82	0
n_overhead_lines	5.665	5.382	10.781	18.433	23.193	15.442	8.4	34.539	44.703	28	3.984	10.331	19.482	25.688	5.761	25.895	45.941	17.74	0.221	1.151	1.751	5.863	249
n_underground_lines	0	0	0.002	0	0.006	0.032	12.3	0.008	0.047	2	0	42.519	63.953	73.67	105.122	139.474	140.92	212.374	2.25	2.049	7	12.95	3
ratio_lines_overhead	1	1	1	1	1	0.999	0.401	1	0.999	0.933	1	0.199	0.24	0.259	0.052	0.179	0.249	0.077	0.074	0.278	0.21	0.353	0.988
min_hops_to_sub	1.002	1.008	1.007	1.001	1.001	1.005	2	1.002	1.014	2	1.051	2	2.009	2	2	2.015	2	2	2.023	2.029	2.02	2.01	3
max_hops_to_sub	4.719	4.827	8.116	10.992	12.752	10.108	9	15.522	19.62	11	4.481	8.857	9.591	11.071	11.693	13.113	17.014	16.61	2.712	3.101	4.541	6.601	36.5
mean_hops_to_sub	3.01	3.099	4.866	6.54	7.535	5.956	5.158	9.116	11.085	5.9	2.812	5.208	5.767	6.364	6.604	7.325	8.971	8.684	2.363	2.585	3.197	4.127	20.336
median_hops_to_sub	3.098	3.253	5.047	6.778	7.776	6.137	5.2	9.383	11.324	5.5	2.85	5.143	5.754	6.298	6.481	7.184	8.757	8.402	2.367	2.614	3.187	4.083	20.833
min_dist_to_sub	1	1	1.008	1	1.002	1.101	31.2	1.035	1.297	23	1.376	17.985	14.995	14.447	19.641	17.781	14.637	18.798	31.81	98.922	30.225	38.294	27.367
max_dist_to_sub	104.6	107.7	214.4	248.8	286.6	383.5	332.7	377.8	526.6	431	77.26	315.69	313.04	367.23	419.63	444.32	540.46	593.36	64.76	163.93	192.98	332.99	1656.8
mean_dist_to_sub	57.63	59.08	109.9	141.7	147.7	215.7	166.7	200.8	277	185.533	35.57	155.99	162.27	182.21	215.33	226.07	262.02	293.09	47.86	132.35	106.75	174.22	805.17
median_dist_to_sub	59.9	61.5	108.9	141.9	146.4	223.8	158.4	202.6	278.46	170	31.05	152.1	161.43	179.91	213.3	221.56	256.6	283.68	47.42	133.7	102.7	169.5	834.21
num_customers	89.806	30.177	4.944	29.99	8.406	4.295	0	8.976	9.977	0	1.907	0	0	0	0	0	0	0	0	0	0	0	0
avg_degree	1.977	1.936	1.805	1.946	1.895	1.825	1.975	1.91	1.923	1.983	1.635	1.975	1.985	1.988	1.988	1.992	1.993	1.993	1.679	1.711	1.843	1.914	1.995
diameter	4.795	4.734	6.633	9.358	11.08	8.229	14.4	13.46	17.1	18	3.9	14.7	16.69	19.19	19.83	21.82	27.19	24.8	4.63	5.108	7.226	10.69	44.33
avg_path_len	2.257	2.385	3.061	3.79	4.675	3.609	6.416	5.563	6.773	8.089	2.04	6.766	7.794	8.725	9.056	9.918	11.959	11.189	2.321	2.517	3.474	4.894	17.926
max_sub_node_distance_mi	0.051	0.051	0.084	0.119	0.139	0.168	0.202	0.183	0.254	0.268	0.019	0.192	0.194	0.225	0.256	0.266	0.324	0.32	0.035	0.096	0.113	0.2	1.014
lv_1ph_len_mi	0	0	0	0	0	0.001	0	0	0.001	0	0.001	0.531	0.291	0.881	0.982	0.593	1.543	1.743	0.045	0.121	0.183	0.379	0
lv_2ph_len_mi	0	0	0	0	0	0	0	0	0.001	0	0	0.039	0.026	0.037	0.767	0.104	0.08	1.461	0.002	0.008	0.006	0.014	0
lv_3ph_len_mi	0.082	0.079	0.179	0.232	0.304	0.35	0.533	0.47	0.644	0.7	0.052	0.344	0.911	0.578	0.117	1.71	1.104	0.373	0.008	0.016	0.048	0.135	6.058
lv_oh_3ph_len_mi	0.082	0.079	0.179	0.232	0.304	0.35	0.194	0.47	0.644	0.663	0.052	0.186	0.366	0.344	0.027	0.549	0.435	0.09	0.001	0.007	0.017	0.075	5.886
num_lv_1ph_loads	0.386	0.418	0.067	0.01	0.009	0.02	61	0.031	0.07	87	0.065	32.278	27.099	66.426	69.03	52.594	115.448	113.293	1.436	1.676	4.058	8.096	0
num_lv_3ph_loads	0.002	0.032	0.007	0.022	0.002	0.013	0.1	0.006	0.107	0	0.017	2.042	34.547	2.82	0.922	62.947	5.309	7.358	0.136	0.108	0.261	0.532	136
avg_num_load_per_transformer	0.308	0.377	0.068	0.031	0.011	0.033	58.3	0.036	0.189	87	0.075	33.995	61.535	68.873	77.316	116.729	120.559	143.813	1.26	1.512	4.028	8.376	136
n_2_wire_lines	0	0	0.003	0	0	0	0	0.015	0.029	0	0.011	2.015	1.617	2.003	37.394	5.902	4.42	77.65	0.083	0.162	0.178	0.486	0
n_3_wire_lines	5.664	5.365	10.746	18.431	23.194	15.448	20.7	34.526	44.647	30	3.929	15.367	56.053	27.851	5.69	104.293	62.74	21.431	0.371	0.334	1.799	4.509	252
n_lines_240V	0	0	0	0	0	0	20.7	0	0	30	0	0	0	0	0	0	0	0	0	0	0.001	0.002	0
n_nodes_240V	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
n_nodes_400V	6.66	6.316	11.674	19.418	24.179	16.376	0	35.489	45.363	0	4.798	0	0	0	0	0	0	0	0	0	0	0	0

Table 10 Relative importance of features to the final clustering (higher value means more important, values sum to 1.0)

More important		Less important	
Feature	**Average Importance**	**Feature**	**Average Importance**
n_nodes	0.096	**max_dist_to_sub**	0.013
num_customers	0.069	**n_3_wire_lines**	0.011
mean_dist_to_sub	0.058	**lv_1ph_len_mi**	0.010
num_lv_3ph_loads	0.053	**n_overhead_lines**	0.009
n_nodes_400V	0.049	**n_deg_1_node**	0.009
mean_hops_to_sub	0.048	**num_lv_1ph_loads**	0.009
n_lines_240V	0.046	**avg_degree**	0.009
total_line_length	0.040	**Diameter**	0.007
lv_oh_3ph_len_mi	0.039	**n_Line**	0.007
min_dist_to_sub	0.038	**n_Load**	0.007
n_1_wire_lines	0.037	**ratio_lines_overhead**	0.004
n_nodes_240V	0.036	**max_hops_to_sub**	0.004
n_2_wire_lines	0.034	**n_deg_2_node**	0.003
lv_2ph_len_mi	0.03	**n_deg_3_node**	0.003
n_PhaseLoad	0.031	**median_hops_to_sub**	0.003
lv_3ph_len_mi	0.030	**n_deg_>4_node**	0.002
max_sub_node_distance_mi	0.030	**n_deg_4_node**	0.002
avg_path_len	0.029	**min_hops_to_sub**	0.001
n_Wire	0.023		
n_Node	0.020		
avg_num_load_per_transformer	0.018		
median_dist_to_sub	0.015		
n_underground_lines	0.015		

4.3 Feature selection comparison

We briefly compare the first 23 features in terms of clustering importance for the LVFT with features identified as important from two other recent network clustering studies (Rigoni V. , Ochoa, Chicco, Navarro-Espinosa, & Gozel, 2016) and (Ma, et al., 2019). In Table 1 we have identified features that are similar (or would be expected to be highly correlated) and highlighted them in the same colours. It is evident that there is not a large amount of overlap between any pair of studies, with much fewer than half of the features selected appearing in common, and only one feature: the number of customers, making an appearance in all three. This suggests that there is more to understand about which features are more relevant for which particular purposes (noting that Ma et a. (2019) selected their features based on explanatory power for estimating losses, whereas neither this project nor Rigoni et al. (2016) singled out any particular features for greater weight in the categorisation). Without further investigation it is not clear the extent to which the inclusion of a feature in the selected classification set is an artefact of the data availability, due to the unique characteristics of the set of networks that were clustered, the unit of analysis, or is likely to be a consistently useful classification feature in general.

Table 11 Comparison of selected features with other clustering studies

LVFT	(Rigoni V. , Ochoa, Chicco, Navarro-Espinosa, & Gozel, 2016)	(Ma, et al., 2019)
Number of nodes	1. Total number of customers (PC 1 to PC8)	1. Median of line loading in LC [%]
Number of customers	2. No. of Domestic Economy 7 Two Rate customers (PC2)	2. No. of customers with power measured devices
Mean distance of all nodes to substation (transformer)	3. No. of Non-Domestic Unrestricted customers (PC3)	3. Relative loss of the grid in LC [%]
Number of 3 phase loads	4. No. of Non-Domestic Unrestricted (PC3) and Economy 7 Two Rate (PC4) customers	4. Median of node degrees of distribution cabinets
Number 400V nodes	5. No. of Non-Domestic Maximum Demand customers (PC5 to PC8)	5. Max. distance between distribution cabinets [km]
mean number of edges from all nodes to substation (transformer)	6. Total conductor length [m]	6. Max. length of lines [km]
Number 240V lines	7. Main path distance [m]	7. Median of line lengths [km]
Total line length	8. Average path impedance [ohms]	8. Total annual demand of loads [kWh]
Low-voltage overhead 3 phase line length (miles)	9. Total path impedance [ohms]	9. Median of rated power per load [kW]
Minimum distance from node to substation (transformer)	10. Daily mean neutral current [A]	10. Max. rated power per load [kW]
Number of single wire lines	11. Mean 3 phase daily active power [kW]	11. Total load positioning factors [kW·km]
Number of 240V nodes	12. Daily mean standard deviation of active power [kW]	12. DG Penetration level w.r.t. transformer rating
Number of 2 wire lines	13. Daily mean standard deviation of 1 phase active power [kW]	13. Max. DG positioning factor [kW·km]
Length of 2 phase LV lines (miles)	14. Mean 3 phase daily reactive power [kvar]	
Number of phase-loads (i.e. phase count weighted number of loads)	15. Power Factor (PF)	
Length of 3 phase LV lines (miles)	16. No. of PV installations	
Maximum distance between node and substation (transformer)	17. PV-supplied demand	
Average path length	18. PV penetration level (n° customers/PV installations)	
Number of wires	19. Mean PV installation capacity	
Number of loads		
Average number of loads per transformer		
Median distance between node and substation (transformer)		

5 Power Flow Cases for the Taxonomy

This section describes the power flow models which result from combining the models of representative low-voltage networks (the network parameters, topology and impedance, only) with demand trajectory data for the load points in each model. These models will be solved for the power flows and the results reported. Although these power flow cases are intended to demonstrate validity of the network models, the simulated currents and voltages may not be representative of Australian networks and the congestions they incur in the real world.

We next describe the datasets that we used to represent customer loads for power flow simulation analysis. Next, we describe performance standards that operational networks should meet. The following section describes the power flow simulation software and presents results of that analysis in terms of the performance standards.

5.1 Customer load data sources

Non-confidential load data is required in order to develop plausible power flow analysis scenarios. There are not many comprehensive load data time series of individual customers that suitably represent Australian energy users, so a search for appropriate data was undertaken. Twenty-three candidate data sets were identified, which are described in Table 18 and Table 19 in Appendix B . Features of each data set include whether the data is PV only, PV and load, or only load; geographic location; number of samples; sampling time resolution; time range; measured characteristics (i.e. real power only, or reactive power, voltage, current and others) and accessibility. A very recent publication, provides a similar data identification exercise and analysis, reporting on 24 data sets of which only a handful are common with those reported here (Kapoor, Sturmberg, & Shaw, est. 2020).

To select suitable data sets for public release of load time series associated with the low-voltage networks, we required at least: a sampling time of 30 minutes or faster, individual customer resolution, data that could be made publicly accessible and data for 300 or more customers. We had a preference for time series that covered at least 12 months of load, with both PV and consumption, were more recent, in Australia, for a broad range of customer sectors and states, and were conveniently accessible with a permissive license.

Ultimately, from the list of identified data sets, the Smart Grid Smart Cities data set[16] proved to be the most suitable. Although this data is from only a single region within Australia (the Ausgrid region in coastal NSW), there are up to 78,000 household customer samples (much fewer with complete data) at 30-minute resolution, for twelve months, with PV production as well consumption disaggregated into both general and controlled (hot-water) load. Furthermore, it is associated with additional demographic data, and corresponds to existing substation scale load data for the same time period. Unfortunately however, this data set does not include commercial

[16] https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Solar-home-electricity-data
https://data.gov.au/data/dataset/smart-grid-smart-city-customer-trial-data
https://webarchive.nla.gov.au/awa/20160615043539/http://www.industry.gov.au/Energy/Programmes/SmartGridSmartCity/Pages/default.aspx

or industrial customers. The Reward-Based Tariff data set,[17] also for residential customers, covers some 500 households in a different geographical region in Australia (Queensland), however it may be encumbered by privacy and confidentiality constraints that make it difficult to release publicly without additional data processing.

For promising data sets identified by Kapoor, et al., (est. 2020), the Pecan Street Dataset[18] appears the most interesting, with data at sampling times of 15 minutes to 1 second, for more than 1000 households, in various locations in the United States, over more than 10 years and covering numerous appliance types. For fast sampling times of individual appliances, the Tracebase data set[19] identified by Kapoor, et al., (est. 2020) appears attractive, with several years of data including some from Sydney locations.

5.2 Demand data description

A subset of the Smart Grids Smart Cities data, the Ausgrid Solar Homes Electricity Data, was analysed in more detail. The following summarises results from an interactive python notebook that studies load and PV data recorded from 2012 June to 2013 May from 300 homes with installed solar (Ratnam, et al., 2017). The original csv data files can be found at the Ausgrid website.[20]

Figure 9 shows a scatter plot of installed PV capacity against average daily load. There is some correlation, albeit weak – the installed capacity is limited to about 7.2 times the average power consumption (0.3 kW/(kWh/day)). It is also possible to observe that PV capacity during the year of data collection was installed in quantised standard sizes, with numerous examples at 1kW, 1.5kW and 2kW. Figure 10 presents a histogram of daily PV generation and daily load, both showing an approximately log-normal relationship, with daily generation typically being less than daily load (that is, net import).

[17] Technical report: Load and solar modelling for the NFTS feeders Part of the Virtual Power Station 2, ARENA-funded project Erin L. Oliver, Cristian Perfumo June 2015, https://data.csiro.au/dap/SupportingAttachment?collectionId=15331&fileId=916

[18] Pecan Street. Dataport: the world's largest energy data resource. Pecan Street Inc, 2015. https://dataport.pecanstreet.org/

[19] Andreas Reinhardt, Paul Baumann, Daniel Burgstahler, Matthias Hollick, Hristo Chonov, Marc Werner, and Ralf Steinmetz. On the accuracy of appliance identification based on distributed load metering data. In 2012 Sustainable Internet and ICT for Sustainability (SustainIT), pages 1–9. IEEE, 2012. https://github.com/areinhardt/tracebase

[20] https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Solar-home-electricity-data

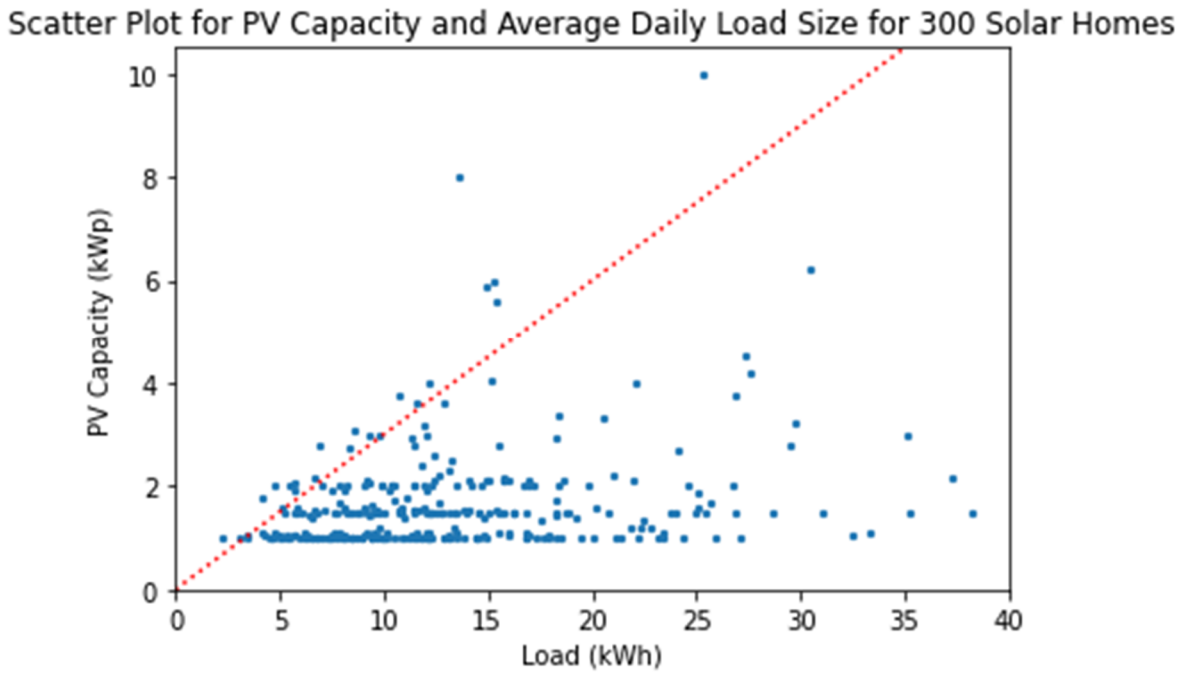


Figure 9 PV Capacity (peak kilowatts: kWp) versus average daily load (kilowatt hours)

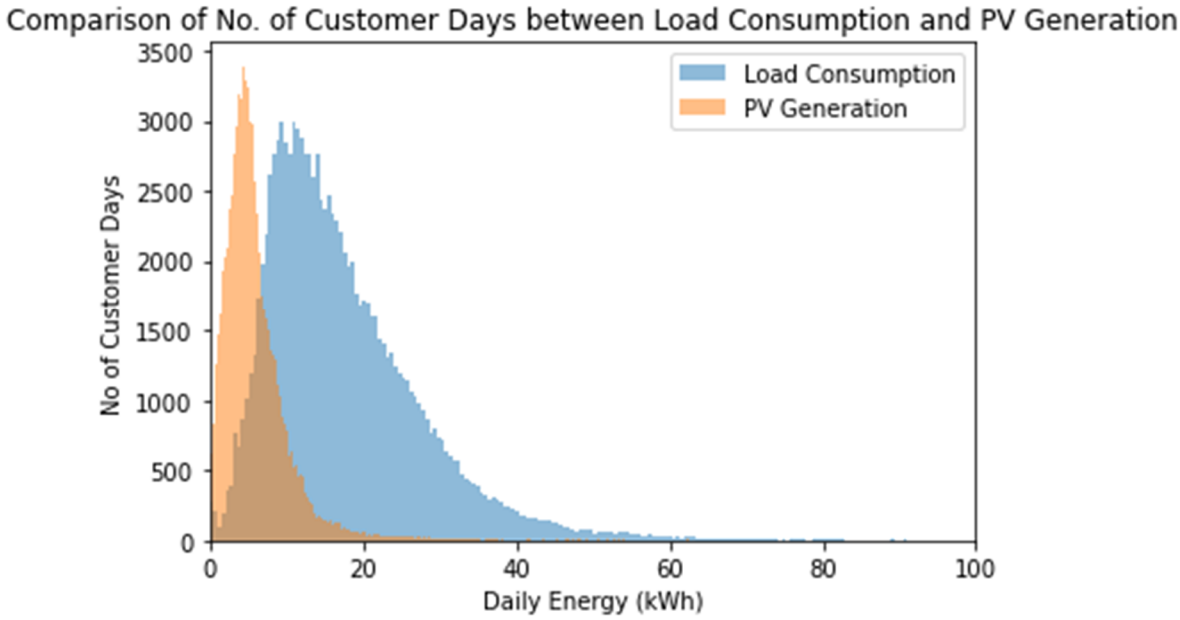


Figure 10 PV generation and load histogram

The diurnal variation of PV generation and load is shown in Figure 11 and Figure 12 with a lower morning peak demand and a higher evening peak (variation by customer is normalised by PV capacity average load respectively). It can be seen by comparing the winter diurnal profile Figure 13 and the summer profile Figure 14 (both normalised by average load over the full year) with the annual profile Figure 12, that the morning peak is much more predominant in Winter, and almost unnoticeable in Summer.

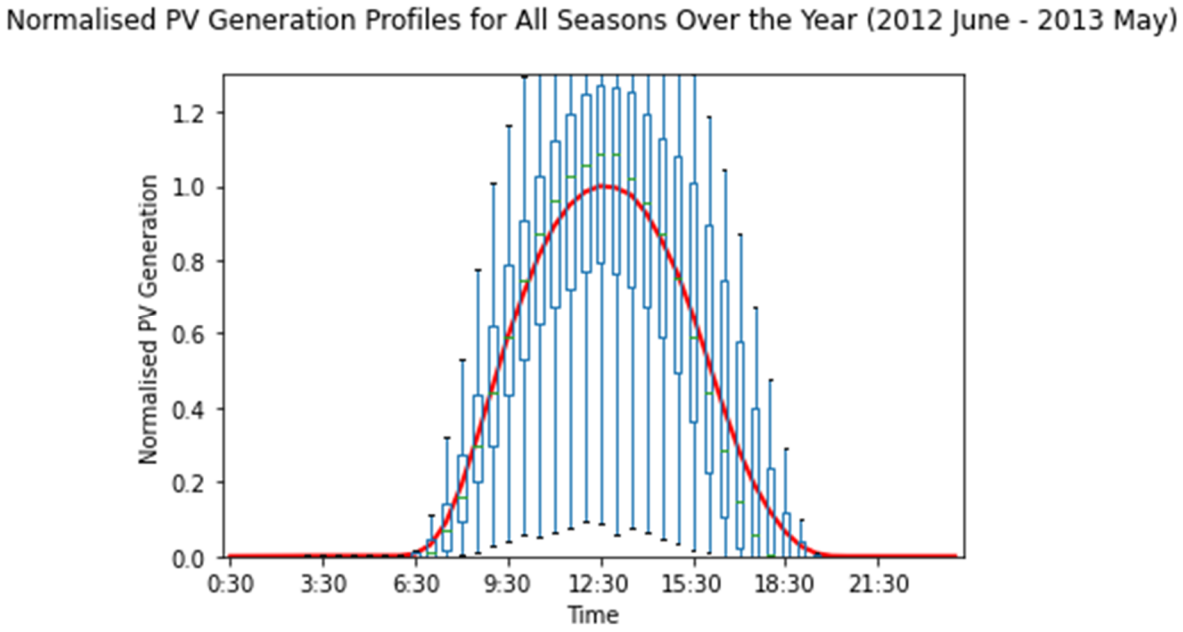


Figure 11 PV generation variation by hour of day (x-axis) and day of year (box-plot)

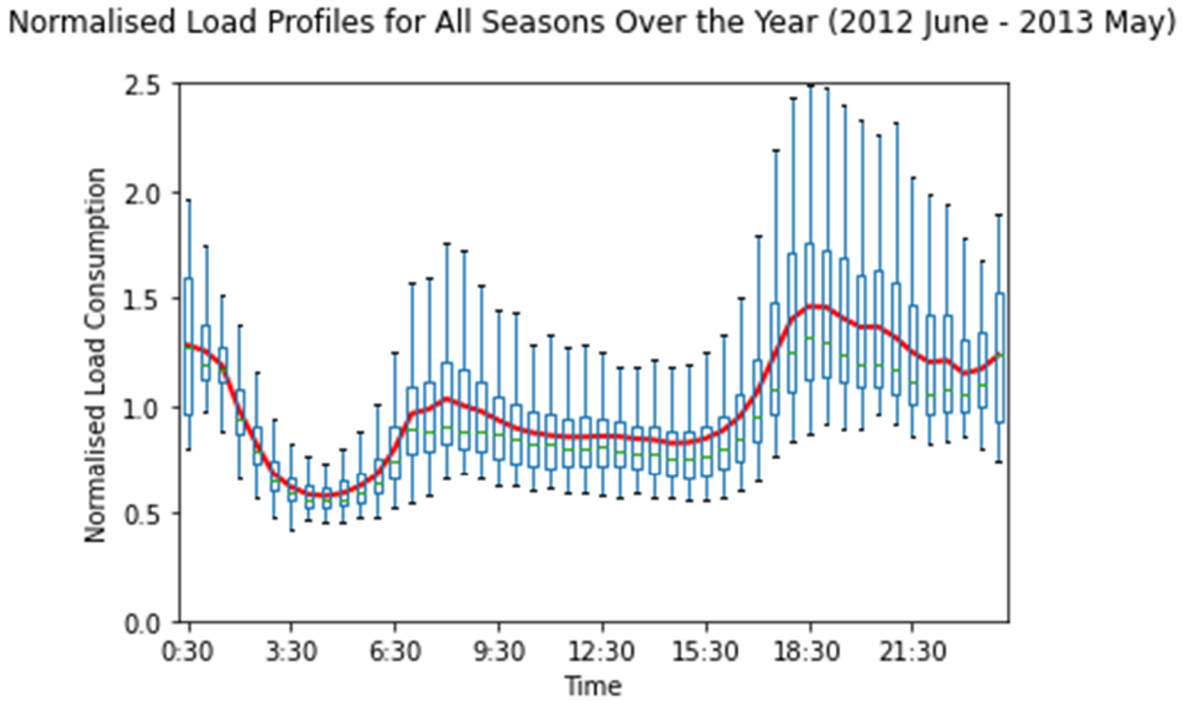


Figure 12 Load variation by hour of day (x-axis) and day of season (box-plot)

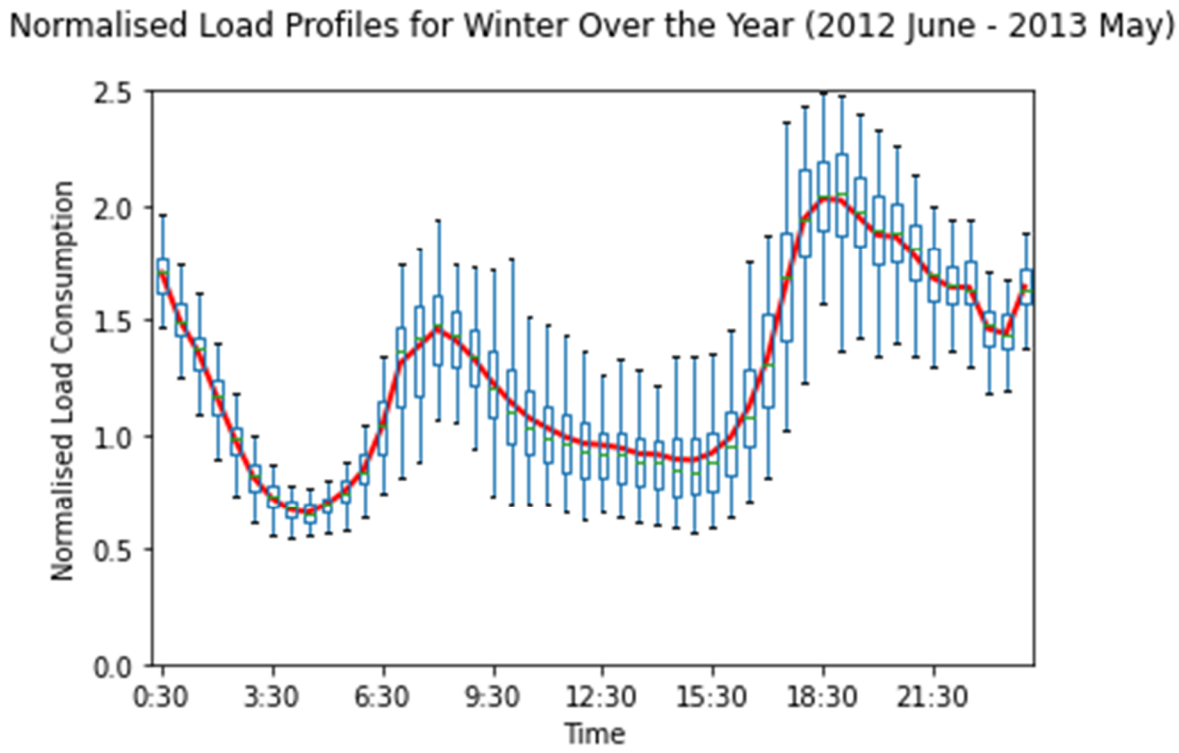


Figure 13 Winter load variation by hour of day (x-axis) and day of year (box-plot)

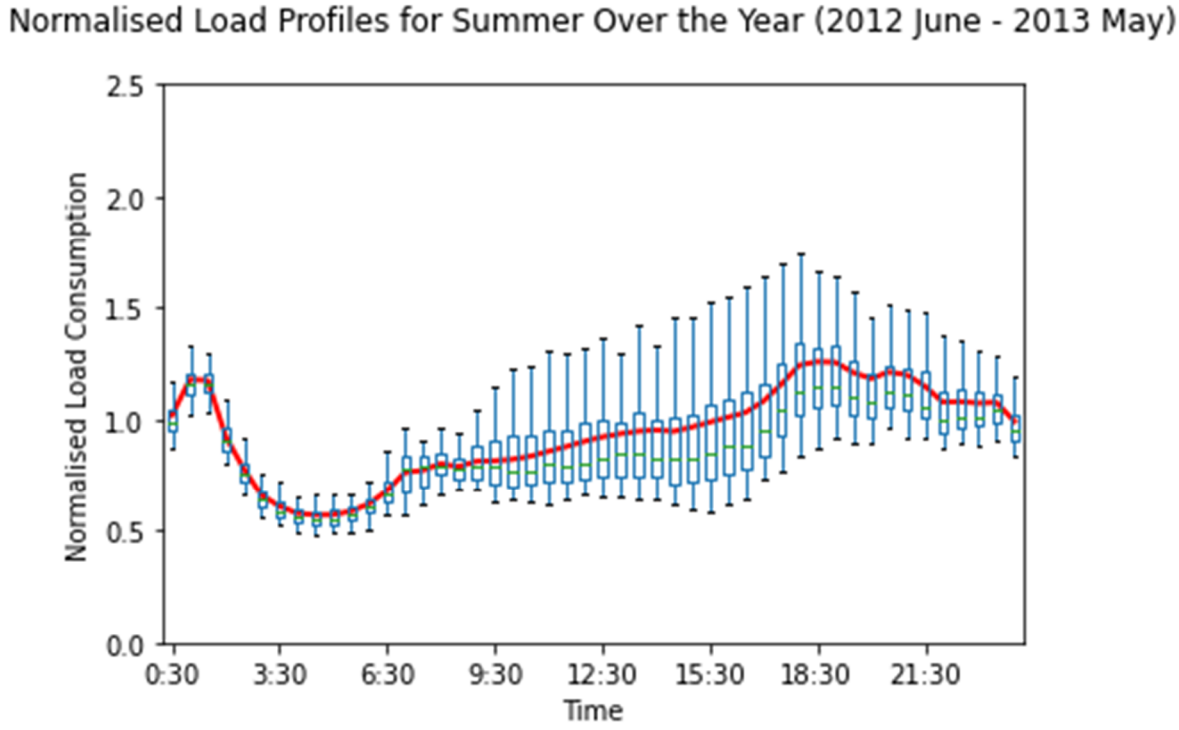


Figure 14 Summer load variation by hour of day (x-axis) and day of season (box-plot)

While the previously discussed figures show daily variation in the diurnal profiles using the box-plots (for the 25th and 75th percentiles), Figure 15 and Figure 16 show variations in the diurnal profile across the customer base (for each of PV generation and load). Note that for PV

generation, the variation across customers is much tighter than variation across days of the year. This is to be expected, as customer variation is due to local shading and daily variation is due to weather and season solar variability. In contrast, the variation of load by customer is slightly greater than variation by day of year (particularly between 1am to 6am). Evidently, individual customers demonstrate repeating patterns of usage from day to day, but different customers show slightly more individually distinctive patterns from household to household.

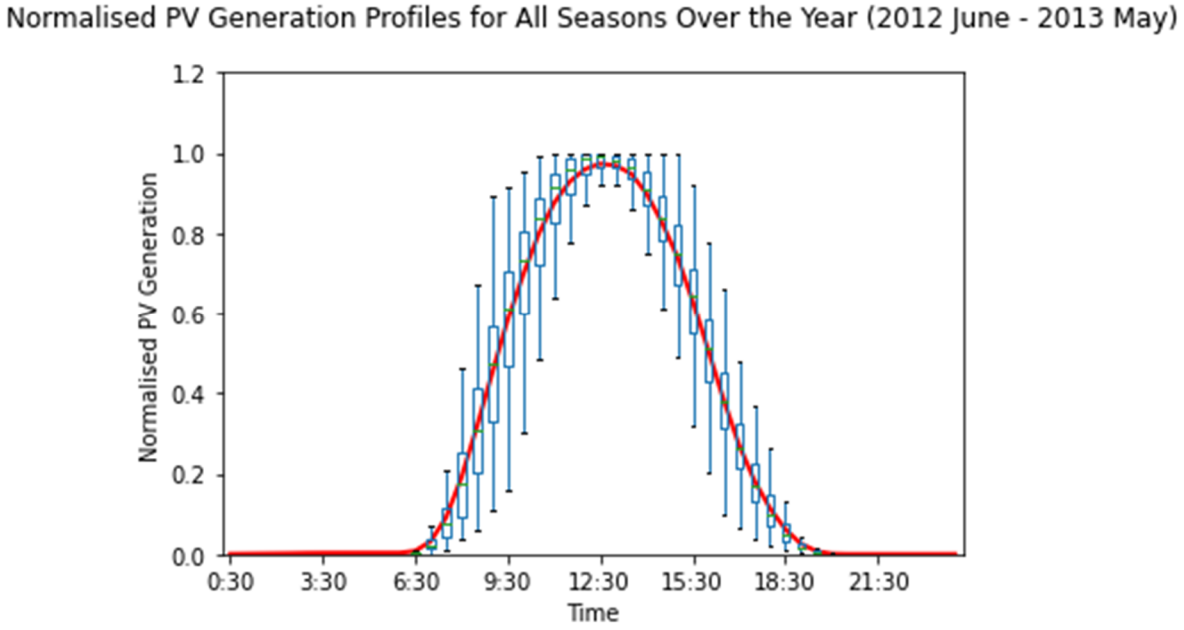


Figure 15 PV generation variation by hour of day (x-axis) and customer (box-plot)

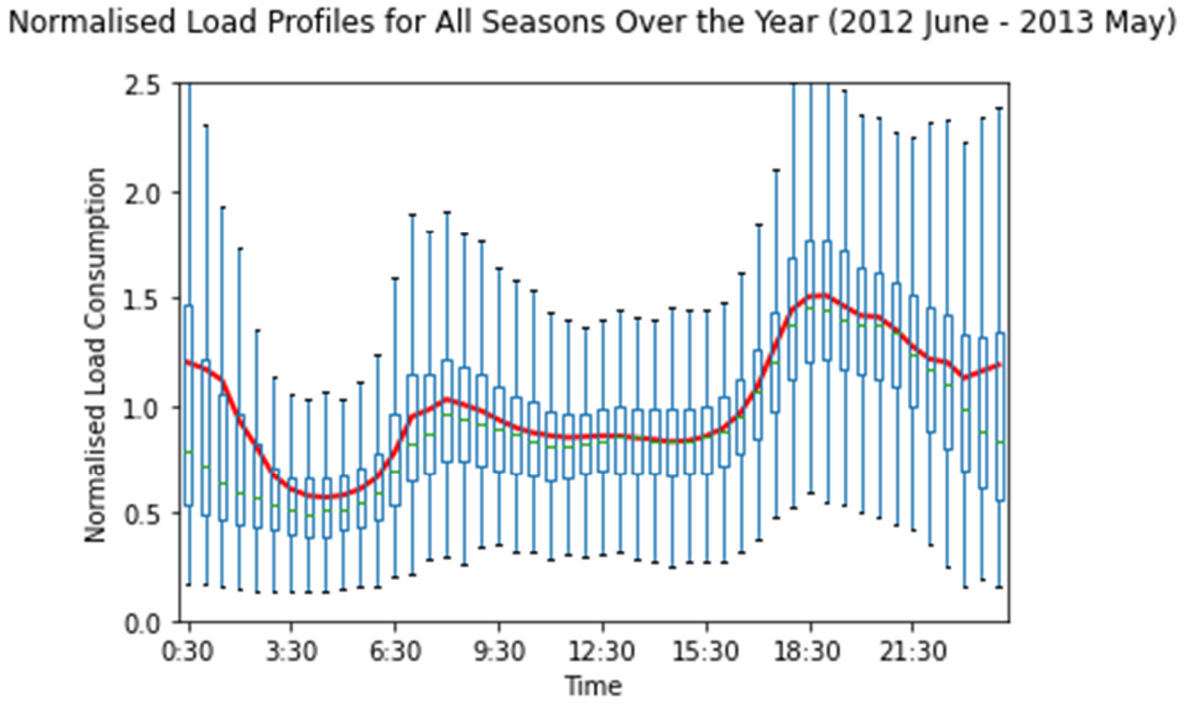


Figure 16 Load variation by hour of day (x-axis) and customer (box-plot)

5.3 Network limits and device data

This section covers performance standards that should be met by operational low-voltage distribution networks in Australia. These include voltage magnitude and harmonic limits,

5.4 Voltage limits

Australian standard 61000.3.100 defines steady-state voltage limits, with values listed in Table 12.

Table 12 AS 61000.3.100 steady state voltage limits

Voltage measure	Phase-to-neutral	Phase-to-phase	Split-phase (1-phase, 3-wire)
1st percentile	216 V	376 V	432 V
99th percentile	253 V	440 V	506 V
50th percentile preferred minimum	225 V	392 V	451 V
50th percentile preferred maximum	244 V	424 V	488 V

Futhermore, the standard defines limits on the 'negative sequence' voltage. The occurrence of negative sequence voltage indicates phase voltage unbalance, and is part of the grid code for this reason. For LV grids the permitted levels are listed in Table 13.

Table 13 negative sequence voltage limits per National Electricity Rules

Non-contingency (30 minute avg)	Credible contingency (30 minute avg)	General (10 minute avg)	Once per hour
2% neg seq.	2% neg seq.	2.5% neg seq.	3.0% neg seq.

After a power flow simulation, performance should be compared to these requirements. We note that Yildiz, et al. (2020) suggest that the occurrence of overvoltage is already common in Australian LV networks, and not just throughout the daytime due to the contribution of PV systems. We note that Australian Standards AS 61000 also specifies limits on harmonics in LV grids. However, it requires additional preparation to set up the data sets necessary to perform harmonic power flow analysis, which we did not set out to do in the LVFT.

5.5 Load and device limits

All devices that are part of the distribution network, or are connected to it, have ratings limits within which they operate safely, reliably and efficiently. Cables, lines and transformers each have individual current and power ratings, as provided by the manufacturers' specifications, and it is preferable that these ratings are not exceeded during network operation. Technologies such as batteries and solar systems have complex sets of parameters that characterise their limits and flexibility.

We adopted OpenDSS' PVSystem and Storage models, and enabled access in the notebooks to a subset of the parameters, see Table 14 for the battery storage parameters and Table 15 for the PV system ones.

Table 14 Exposed battery storage properties

Quantity	Label	Unit	Default value
Energy capacity	kWh rated	kWh	5
Initial energy content	Stored	% of kWh rated	50%
Inverter phases	Phases	Subset of {a,b,c}	{a,b,c}
Inverter rating	kVA	kVA	5
Inverter configuration	Conn	'wye' or 'delta	Wye
Inverter power factor	PF	-	1

Table 15 Exposed PV system properties

Quantity	Label	Unit	Default value
Maximum power point	Pmpp	kW	5.5
Mode	Mode	'constant PF' or 'volt-var'	constant PF
Inverter phases	phases	Subset of {a,b,c}	{a,b,c}
Inverter rating	kVA	kVA	5
Inverter configuration	Conn	'wye' or 'delta	wye
Inverter power factor	PF	-	1

Finally, we model demand response through conservation voltage reduction. The load power consumption is considered to be voltage-sensitive with an exponential model for which the exponent can be chosen. The parameters are indicated in Table 16.

Table 16 Exposed load voltage sensitivity settings

Quantity	Label	Unit	Default value
Active power exponent	cvrwatts	-	0.4
Reactive power exponent	cvrvars	-	0.8

5.6 Power flow validation and notebooks

We developed a set of Julia/Pluto notebooks to enable the use of the low-voltage network models identified in Section 4 for simulation experiments using OpenDSSDirect.jl. OpenDSSDirect is a is a cross-platform Julia software package that implements a 'direct' interface to the OpenDSS power flow engine, using the OpenDSS API in the C programming language. OpenDSSDirect is distributed using the Julia package manager. It is noted that a Python equivalent, i.e. OpenDSSDirect.py, is also available. The OpenDSS low-voltage network files included in our dataset are, by default, snapshots i.e. represent only a single moment in time. We attach time series data in the notebooks programmatically. The load values in the notebooks should not be understood to be

representative of loads actually observed on the networks represented by the models, but merely as a plausible value for a customer load.

The notebooks enable the selection of any of the 23 low-voltage network cluster medoid models using a dropdown list, as in Figure 17.

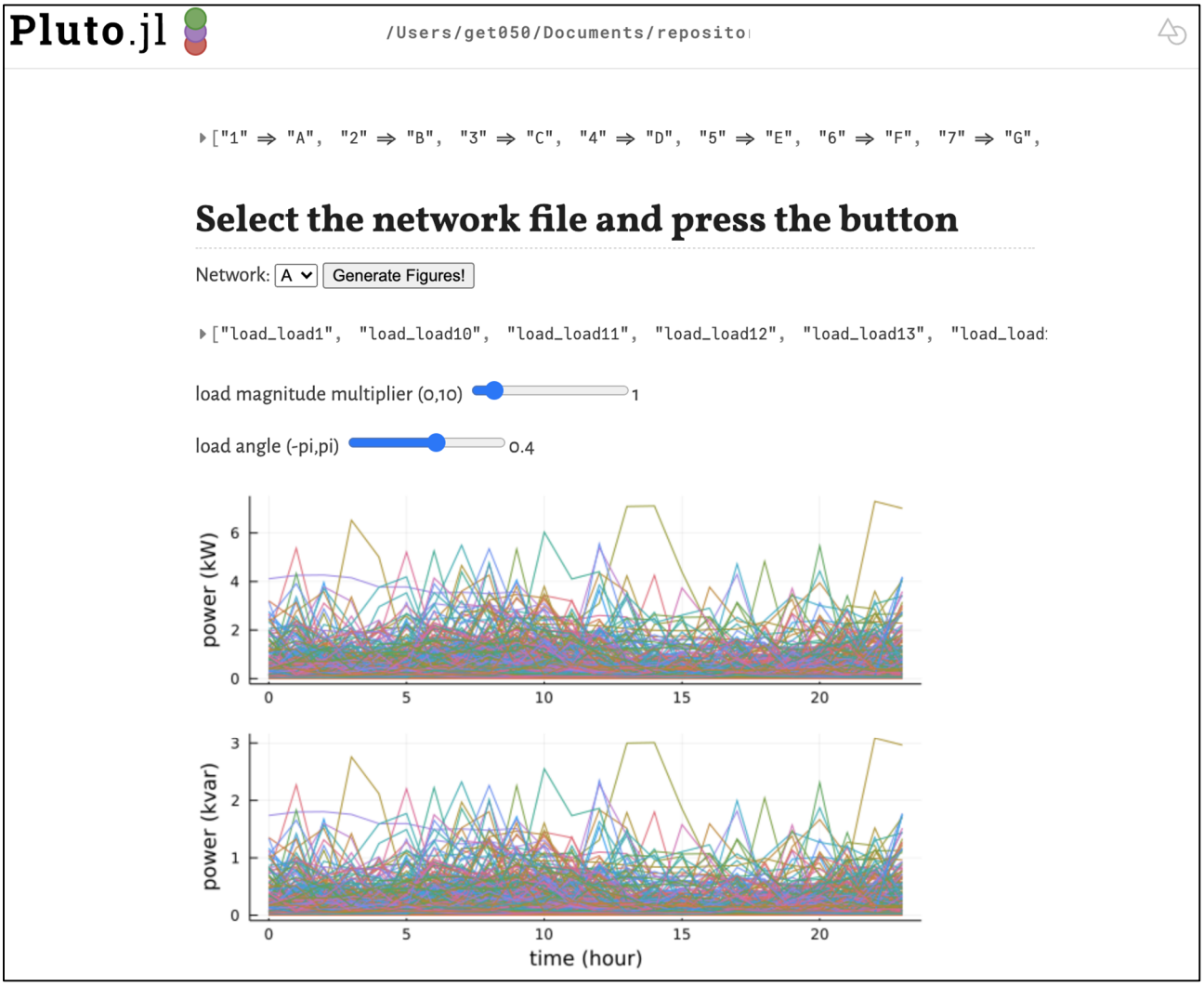


Figure 17 Example Pluto page to select a network in the notebook, scale load power consumption, and attach time series

Next, the notebook user may attach load time series data. In this example, we re-use time series from the Smart Grid Smart Cities project.

Inspect power flow results

We extract information for all transformers, generators, capacitors, lines from OpenDSSDirect and store them in dataframes. We furthermore extract load, bus and pvsystem data to dictionaries.

▸Dict("24" ⇒ Dict("vma" ⇒ [249.62, 249.194, 249.324, 249.379, 249.163, ⋯ more , 249.

Line parameters:

	Bus1	Bus2	Co	C1	CMatrix	EmergAmps	Geometry	Len
7	"09.1.2.3"	"08.1.2.3"	1.6	3.4	-0.6 2.8 -0.6 -0.6 -0.6 2.8	600.0	""	0.0
8	"21.1.2.3"	"09.1.2.3"	1.6	3.4	3×3 Matrix{Float64}: 2.8 -0.6 -0.6 -0.6 2.8 -0.6 -0.6 -0.6 2.8	600.0	""	0.0
9	"14.1.2.3"	"15.1.2.3"	1.6	3.4	3×3 Matrix{Float64}: 2.8 -0.6 -0.6 -0.6 2.8 -0.6 -0.6 -0.6 2.8	600.0	""	0.0
10	"15.1.2.3"	"16.1.2.3"	1.6	3.4	3×3 Matrix{Float64}: 2.8 -0.6 -0.6 -0.6 2.8 -0.6 -0.6 -0.6 2.8	600.0	""	0.0
⋮ more								
18	"23.1.2.3"	"24.1.2.3"	1.6	3.4	3×3 Matrix{Float64}: 2.8 -0.6 -0.6 -0.6 2.8 -0.6 -0.6 -0.6 2.8	600.0	""	0.0

```
lines_df
```

Note that some of the networks don't have transformer data, which will lead to an empty dataframe.

Transformer Parameters:

	IsDelta	MaxTap	MinTap	Name	NumTaps	NumWindings	R	Rneut	⋮ more
1	false	1.1	0.9	"1"	32	2	0.0	-1.0	

```
transformers_df
```

Figure 18 Examples of lines and transformer data in dataframes

Finally, the user can inspect the results of the power flow simulation with the selected loads, and judge them against power quality standards for public networks described in Section 5.4. For instance, voltage magnitudes are subject to:

- phase-neutral voltage limits
- neutral voltage rise limits

- negative sequence voltage magnitude or voltage unbalance factor limits.

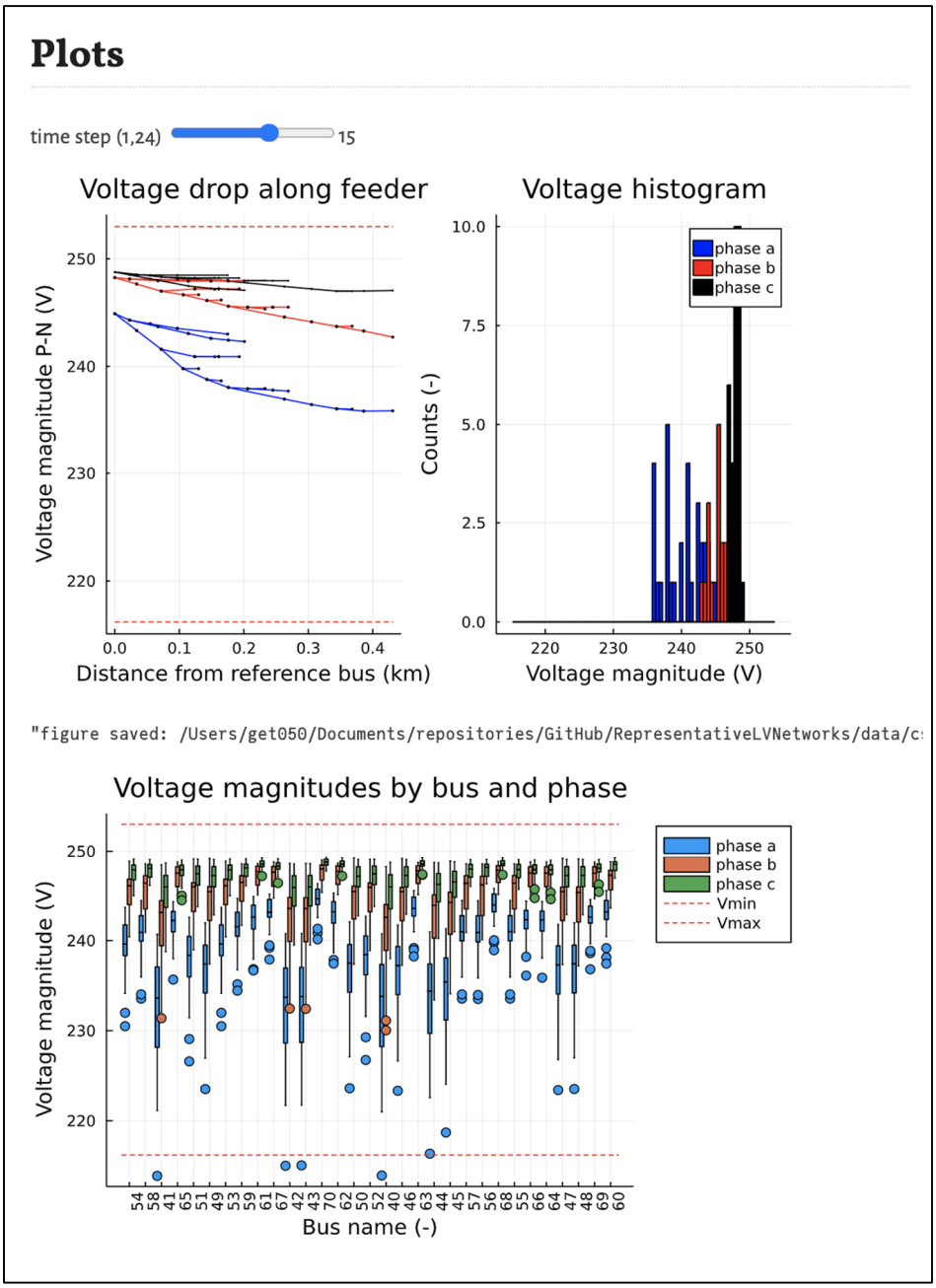


Figure 19 Example of voltage plots in a scenario with significant phase unbalance. The user can drag the slider so see the voltage drop change over time in the top figures.

5.7 Base case power flow simulation results

Partial results of the power flow analysis are presented in this section for the base cases (without PV generation). Results for only a selected sample of the cluster medoid low-voltage network models without PV generation are presented in the main body of this report. The results for the remaining models without PV appear in Appendix C . Simulation results for the 23 selected low-voltage network with distributed energy resources are presented in Section 6. This includes the addition of rooftop solar PV (Section 6.1), a small number of case studies with batteries (Section 6.2) and the full set of networks with conservation voltage reduction (Section 6.3).

5.7.1 Network A

The power flow through the substation bus, for an hourly time series over one day, is visualised first (Recall key features of each LV network cluster appear in Table 7). In this case, the loads are three-phase balanced, so the phase values overlap. Positive values for real and reactive power indicate net consumption on the network downstream, negative values would indicate net injection from the network downstream.

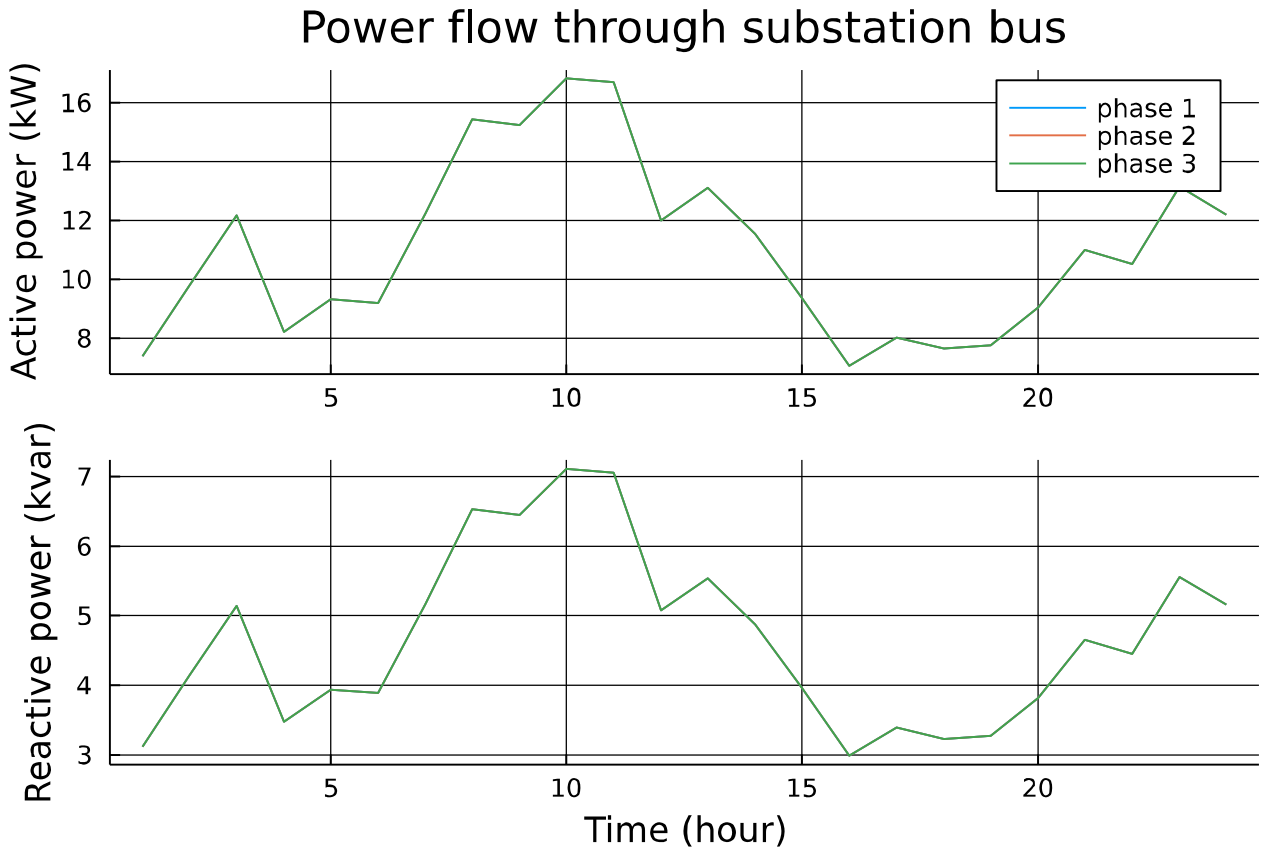


Figure 20 Power flow over time through the substation of network A

The next plot summarises the voltage magnitudes over time (one day) for all the buses, and all phases at each bus, using boxplots. The 0.94 and 1.1 per unit voltage limits, relative to phase-to-neutral voltages are indicated with red lines. The buses are in no particular order.[21]

In this network the range of voltage drops is very much negligible, with voltages in all phases, all times, and all buses, staying close to that imposed by the reference bus, here set at 433 V phase-to-phase (250V phase-to-neutral).

[21] The circular markers on bus 54 represent outliers, though in this case, the chart scale is such that they are not distinguishable from the box plot range.

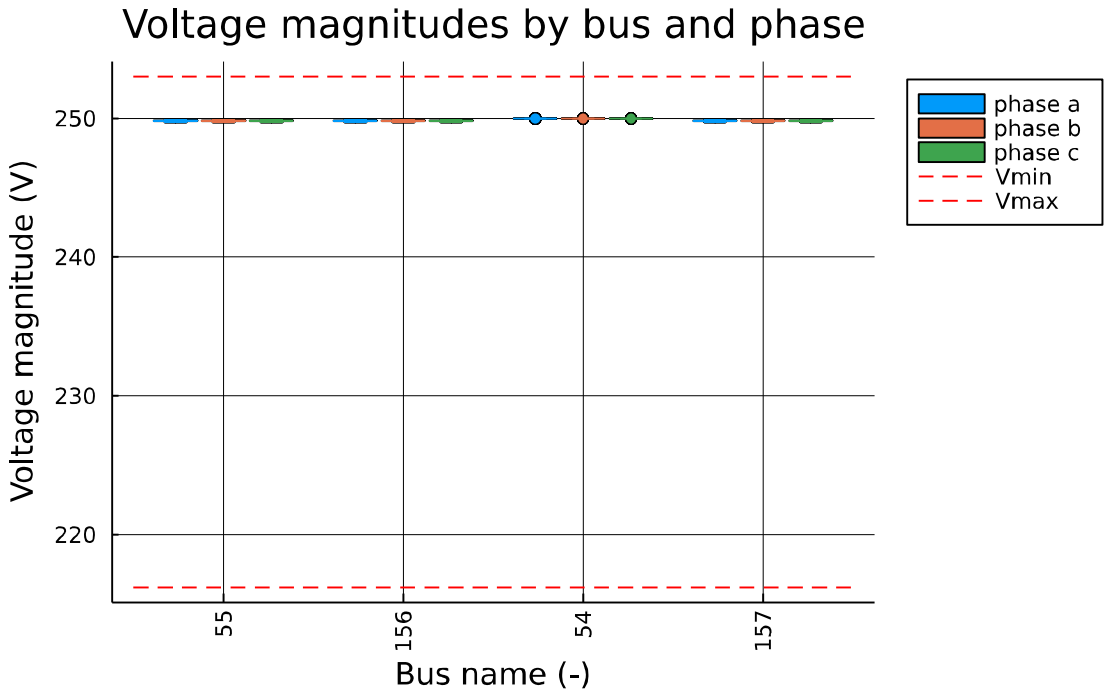


Figure 21 Voltage magnitude boxplots per bus, summarising voltage dynamics in network A

5.7.2 Network D

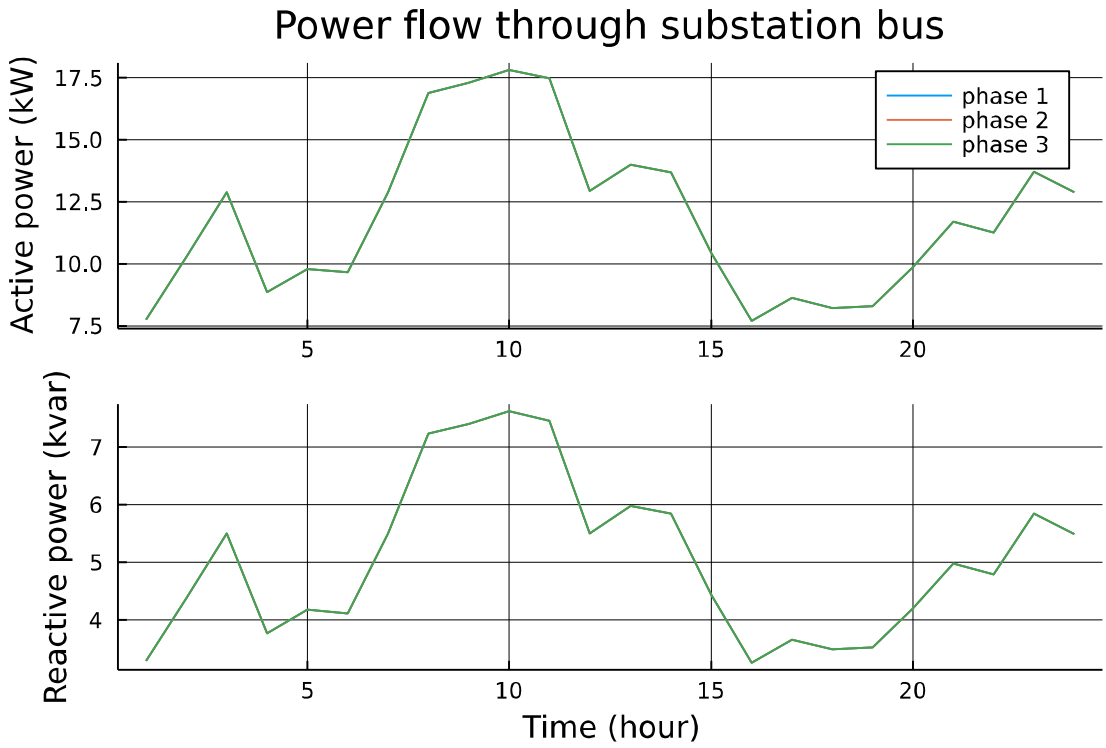


Figure 22 Power flow over time through the substation of network D

In Network D, again with three-phase balanced loads, we see some slight voltage variation over time, and it is now more obvious on this scale that the visualisation depicts boxplots. Because the

load setup is three-phase, the voltage drops in each phase are equal, and the power flow in the transformer is still balanced.

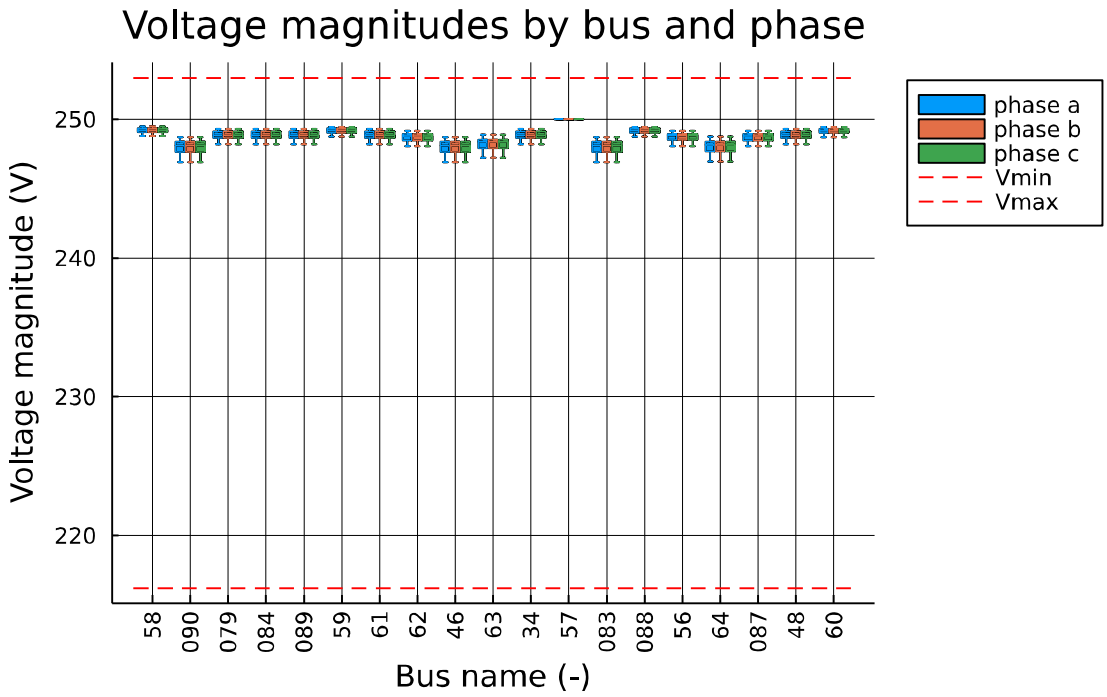


Figure 23 Voltage magnitude boxplots per bus, summarising voltage dynamics in network D

5.7.3 Network G

Network G had phase-specific information, which more easily allows distinct (and unbalanced) loads to be assigned to each phase. The power flow (both real and reactive) through the transformer is now different between the phases (here shown on the primary side).

We furthermore see that phase C (green) has higher voltage deviations across time than the other two phases, for all buses. Nevertheless, the voltages are still very close to the reference of 433 V/sqrt(3) = 250 V. Further analysis shows that unbalance is still very limited overall (<2% negative sequence.

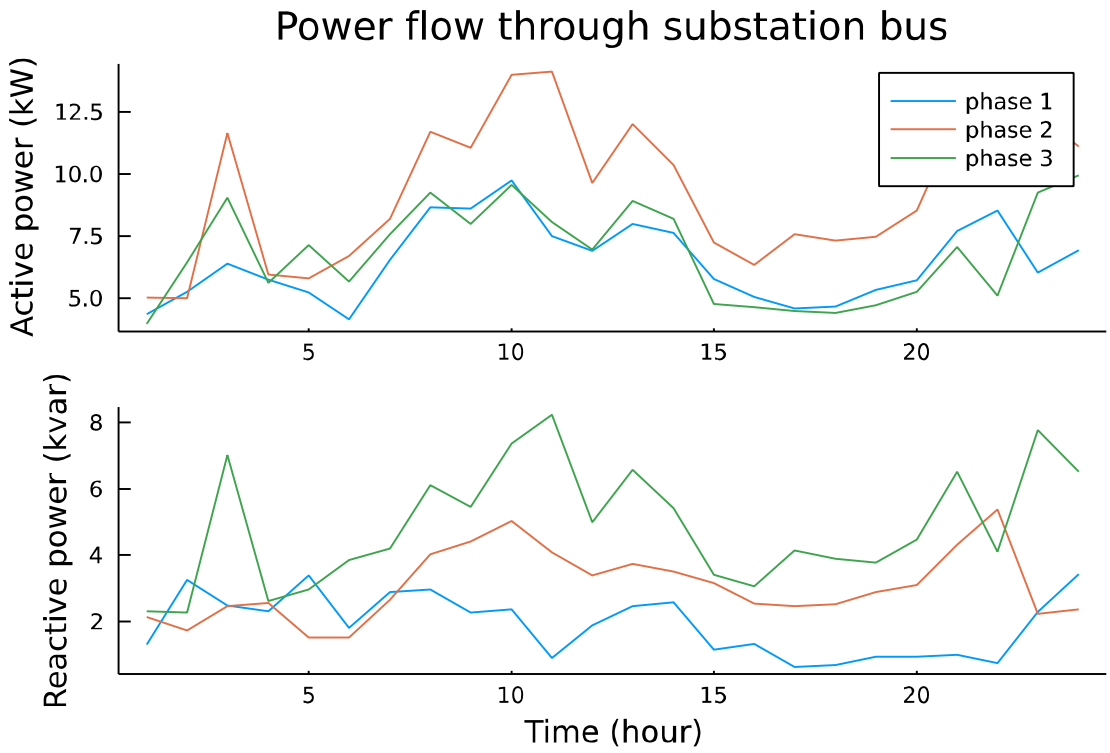


Figure 24 Power flow over time through the substation of network G

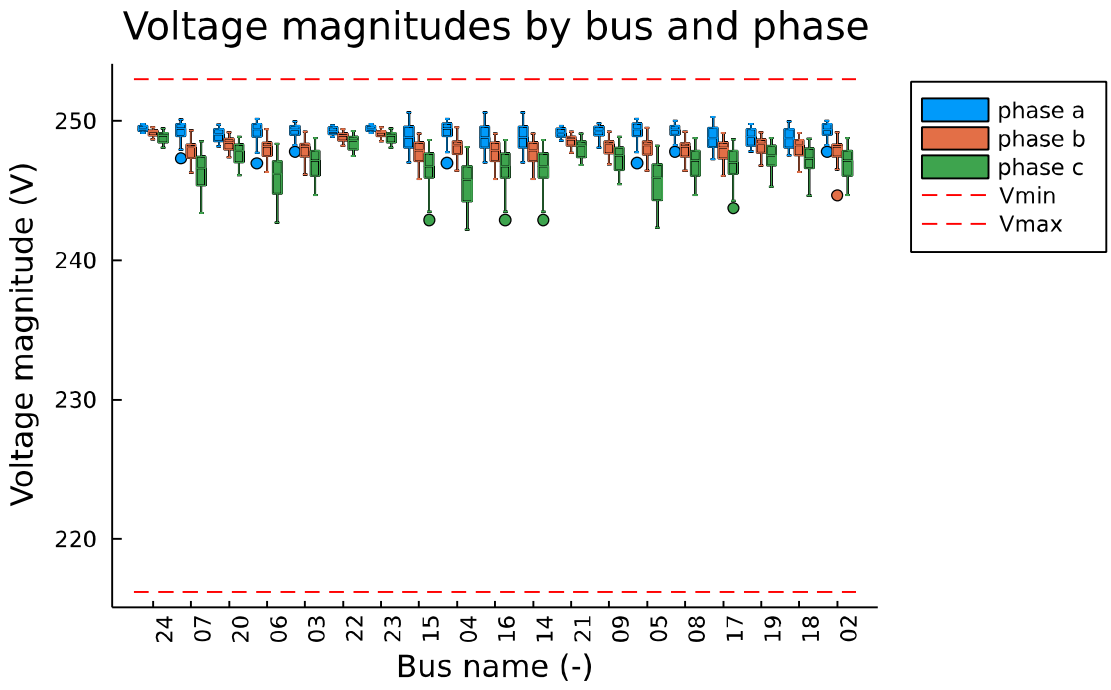


Figure 25 Voltage magnitude boxplots per bus, summarising voltage dynamics in network G

5.7.4 Network J

Network J, again with unbalanced load, has a significant voltage drop in phase A (blue) compared to B and C. Which particular phases suffer the greatest voltage drop, and the range of variation, depends significantly on how time series profiles are assigned to each load

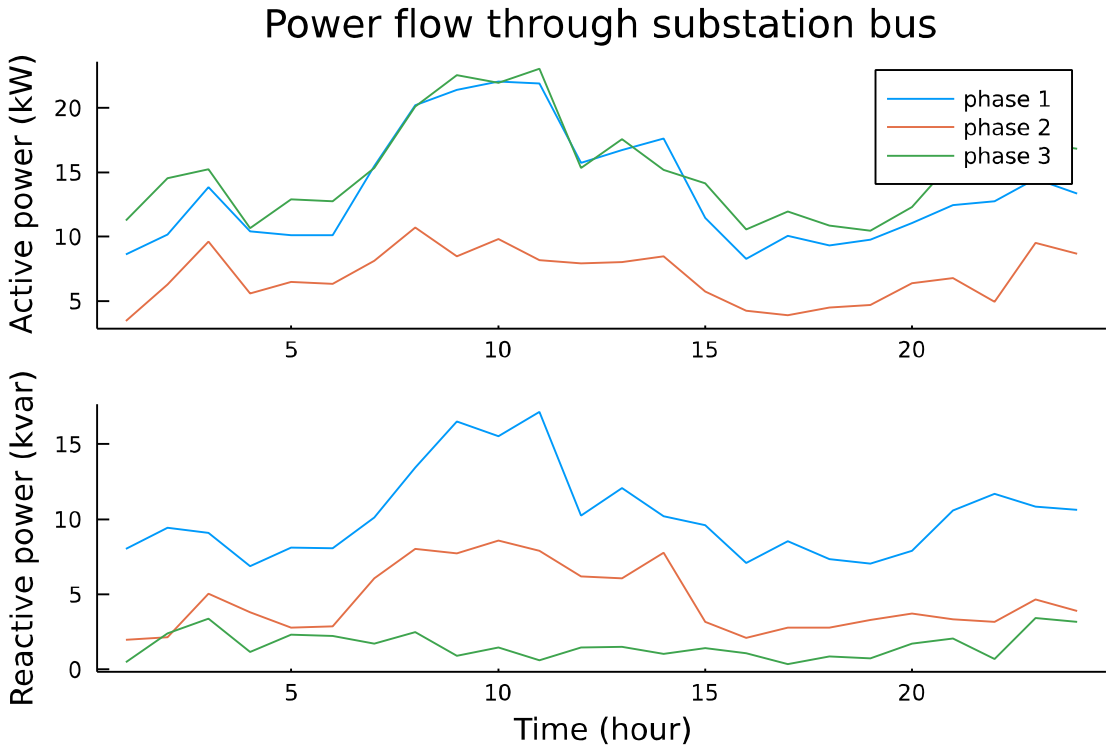


Figure 26 Power flow over time through the substation of network J

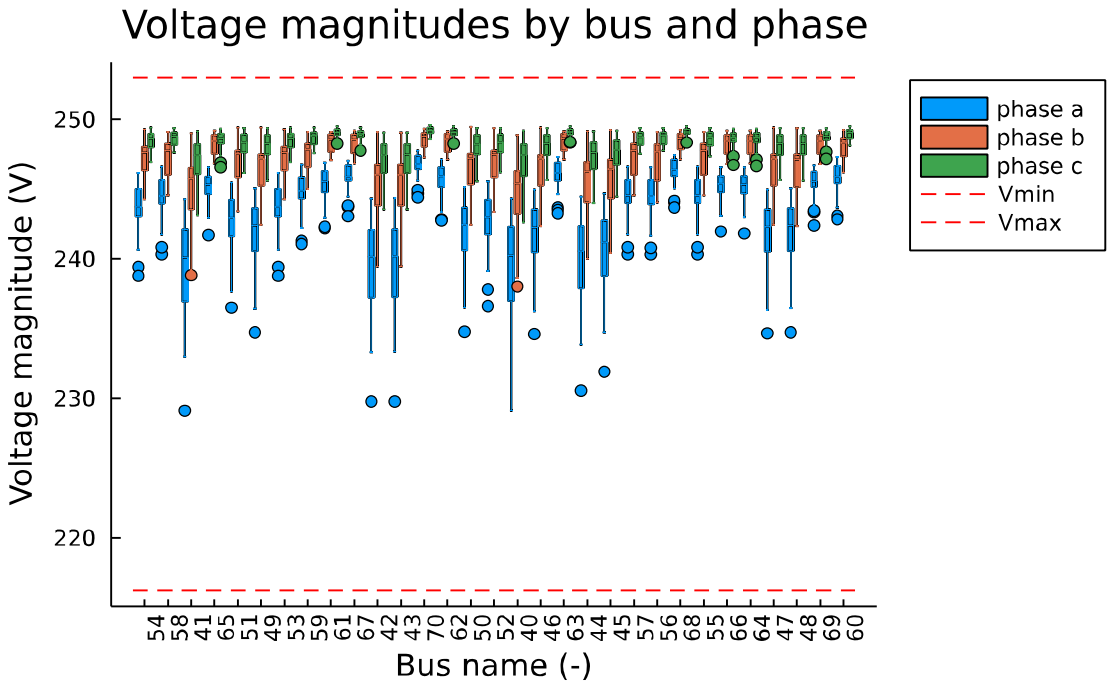


Figure 27 Voltage magnitude boxplots per bus, summarising voltage dynamics in network J

5.7.5 Network L

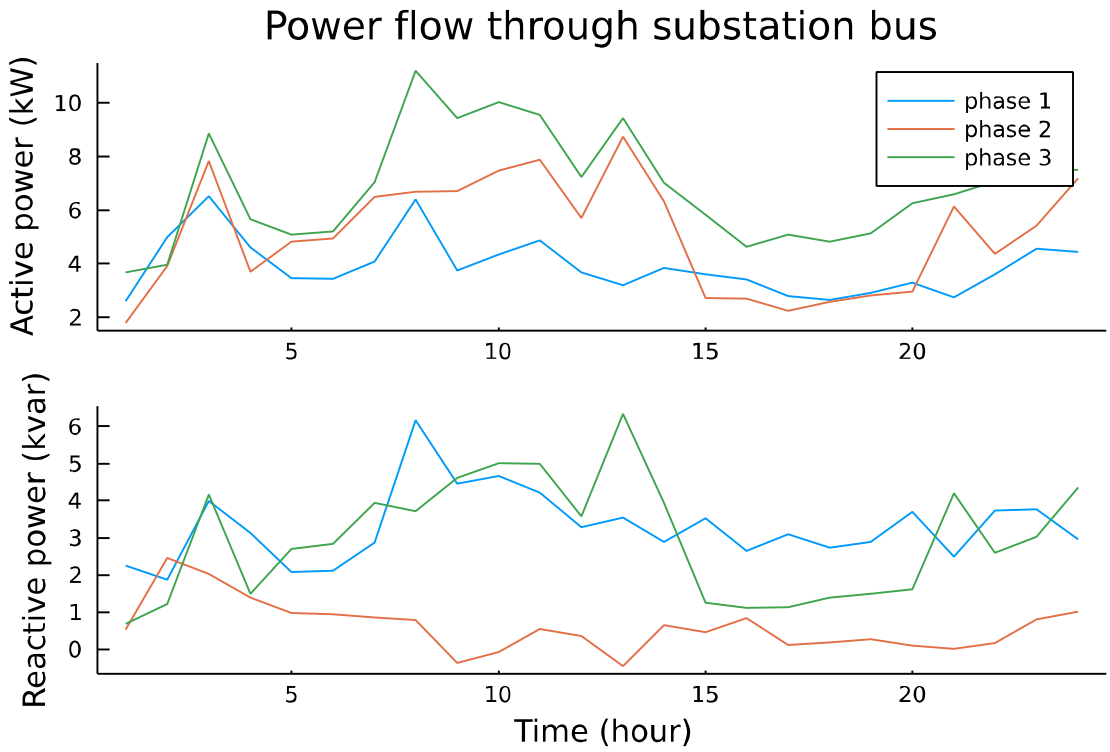


Figure 28 Power flow over time through the substation of network L

In this example for Network L, the voltage variation over time is of moderate magnitude. However, for the selected assignment of load profiles to each phase, the effects among phases evidently largely cancel out, so that none of the three individual phases seems to differ markedly in performance from the other two.

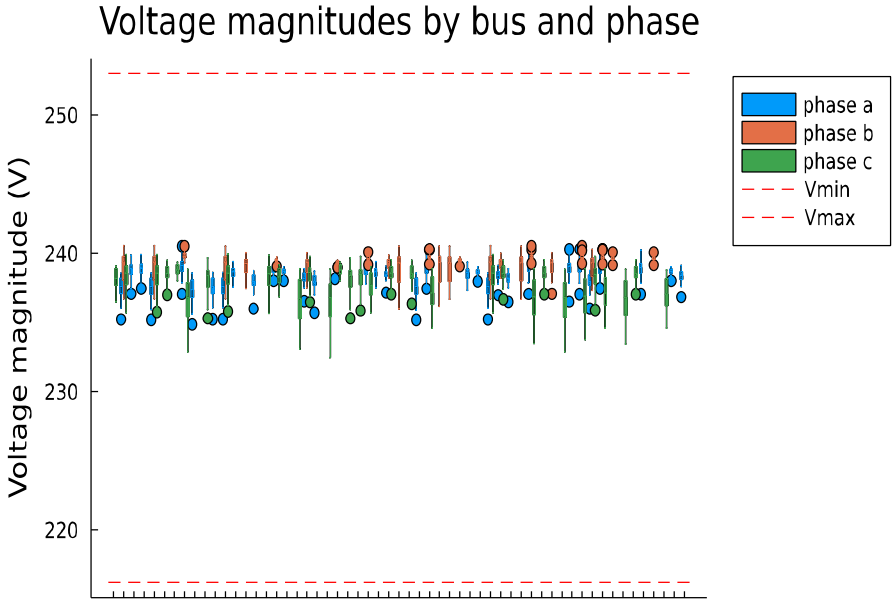


Figure 29 Voltage magnitude boxplots per bus, summarising voltage dynamics in network L

5.7.6 Network Q

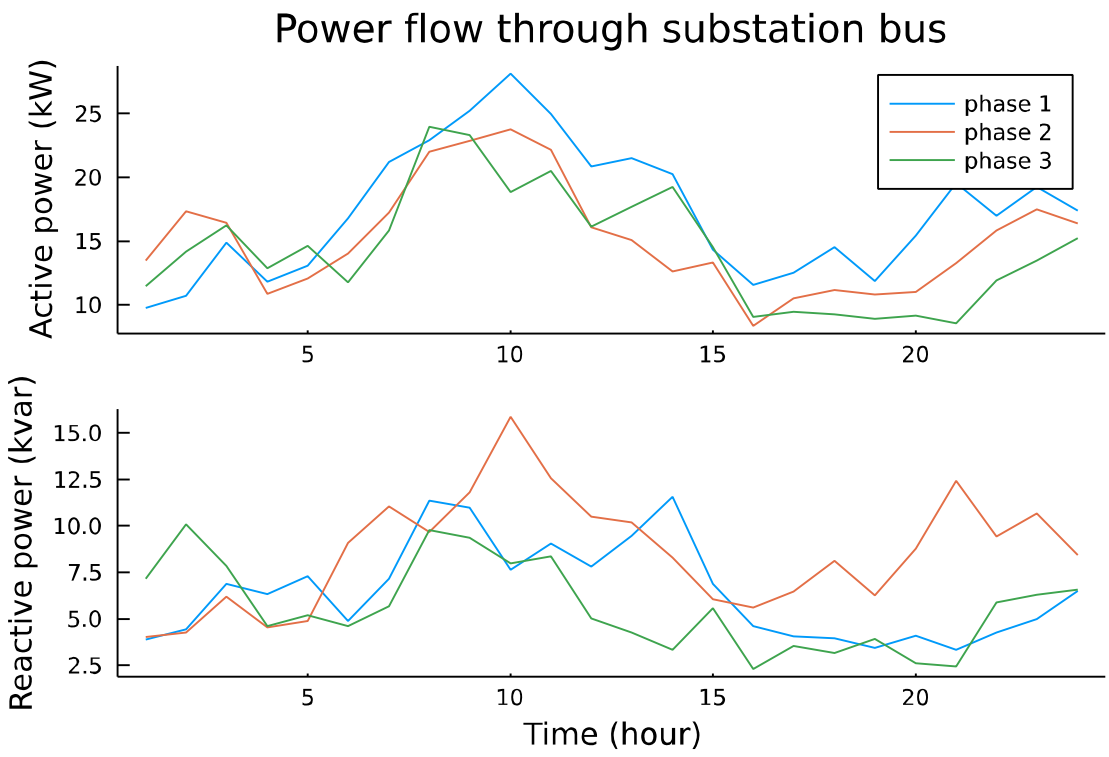


Figure 30 Power flow over time through the substation of network Q

Network Q has both significant phase voltage unbalance and significant voltage drop. We are unsure whether this is due to errors in the network model, perhaps introduced during data parsing, or whether the relatively poor performance is indicative of the challenge of managing the actual network that this model represents

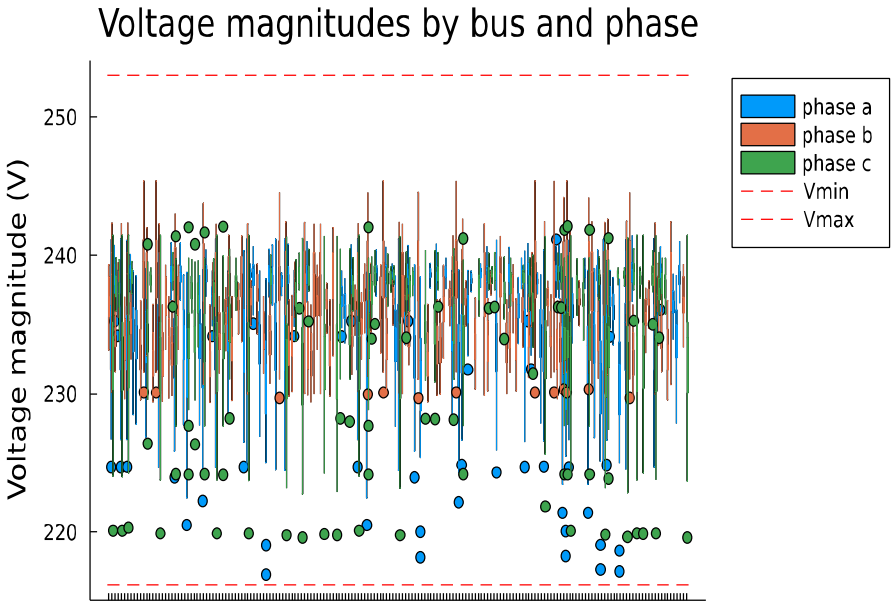


Figure 31 Voltage magnitude boxplots per bus, summarising voltage dynamics in network Q

5.7.7 Network S

Network S has only a sole, single-phase customer. A relatively small magnitude load has been assigned, and there is little variation displayed in voltage over time. Nevertheless, owing to the relatively long length (high impedance) of the feeder, there is a reasonable voltage drop magnitude from the reference bus. In this case we have set the reference bus to 225V (0.9 Vpu) resulting in a load voltage at about 218V, which is outside the target range.

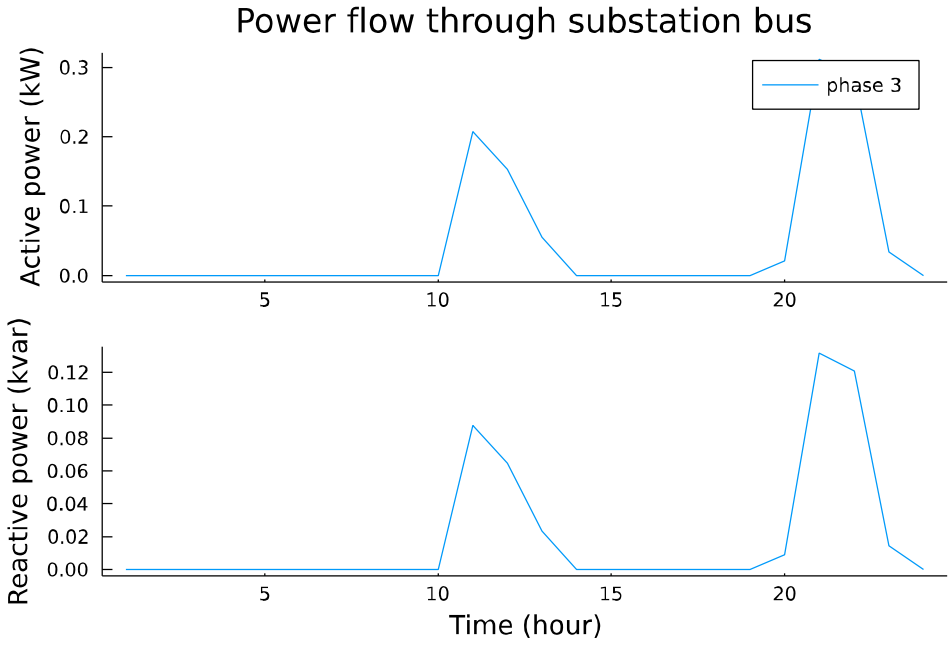


Figure 32 Power flow over time through the substation of network S

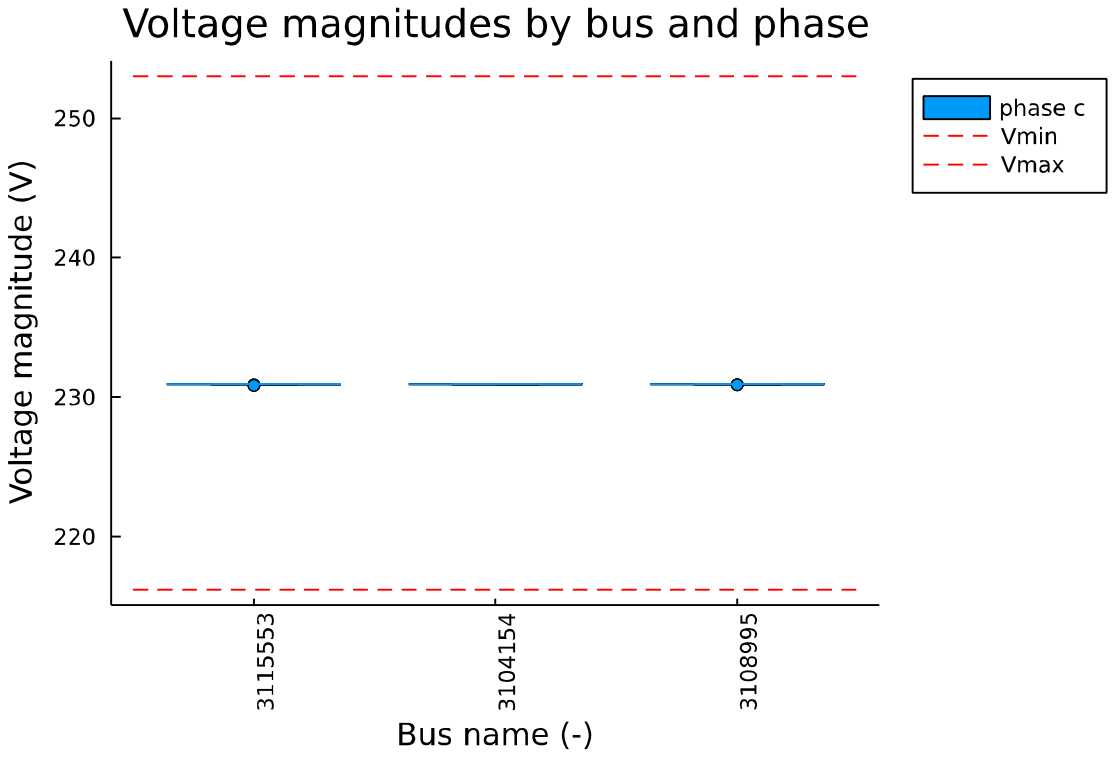


Figure 33 Voltage magnitude boxplots per bus, summarising voltage dynamics in network S

6 Power Flow Cases with Distributed Energy Resources

This section describes power flow models which result from adding solar PV and batteries to network models and load profiles discussed in the previous section. These case studies are solved for the operational power flows. Again, results from only a subset of the 23 low-voltage networks will be presented here, selected to present a range of typical modeled performance.

6.1 Solar PV assignment

The notebook assigns PV systems to the same buses to which loads are connected in the base cases. However, the number of PV systems that are added can be varied by the user. By default, the PV systems are three-phase, but tick box options in the notebooks can be used to restrict all the PV systems to individual phases. The PV inverter rating can be varied, although it must be identical for each PV system. By default, (all) the PV systems are wye connected, but delta connected may alternatively be selected. By default, the PV systems operate in constant power factor mode at 0.95 pu, but it is possible to instead simulate volt-var control as well. We use droop settings as specified for the state of Victoria: the linear range of volt-var behaviour occurs from 0.95 to 1.05 pu voltage, with maximum var injection below 0.95, and maximum absorption above 1.05 pu.

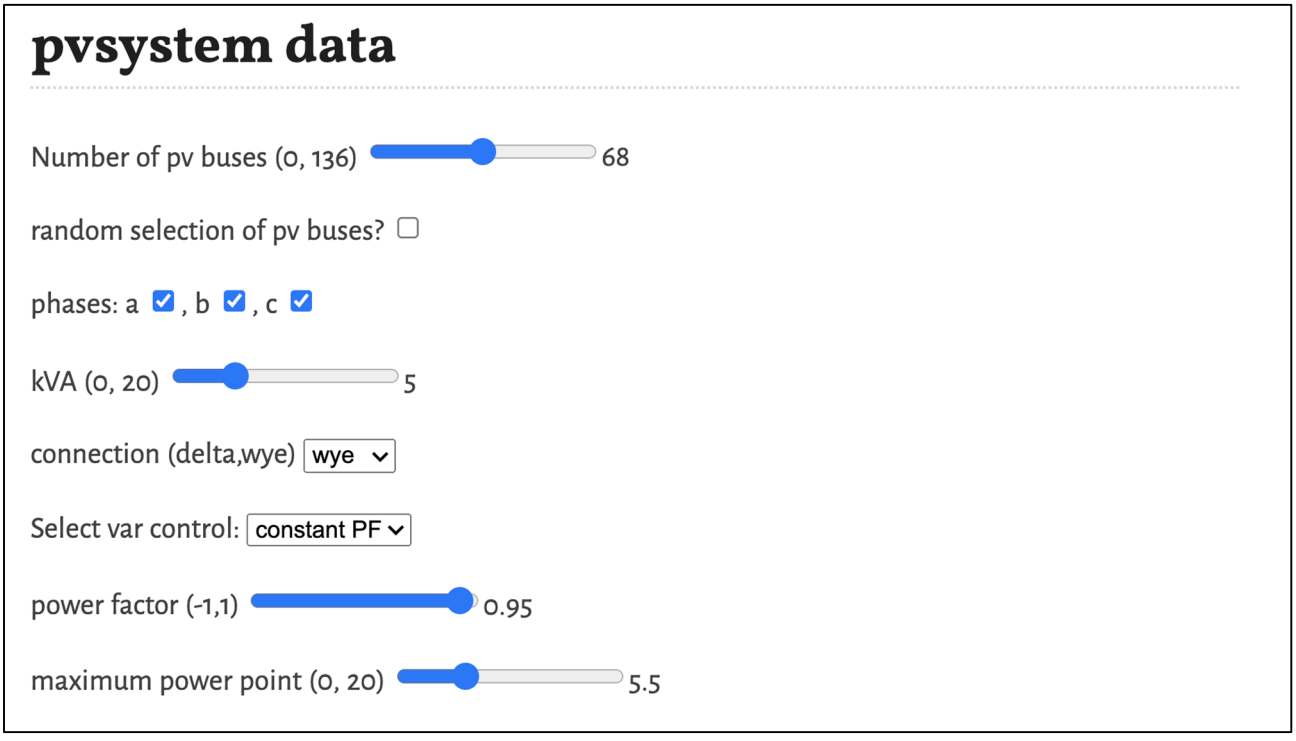


Figure 34 PV system settings that can be tuned

The results presented in this section use the default settings shown above, with 50% of the loads having PV.

6.1.1 Network A

The simulation results for Network A show power flow through the substation becoming negative during the daytime, indicating reverse power flow.

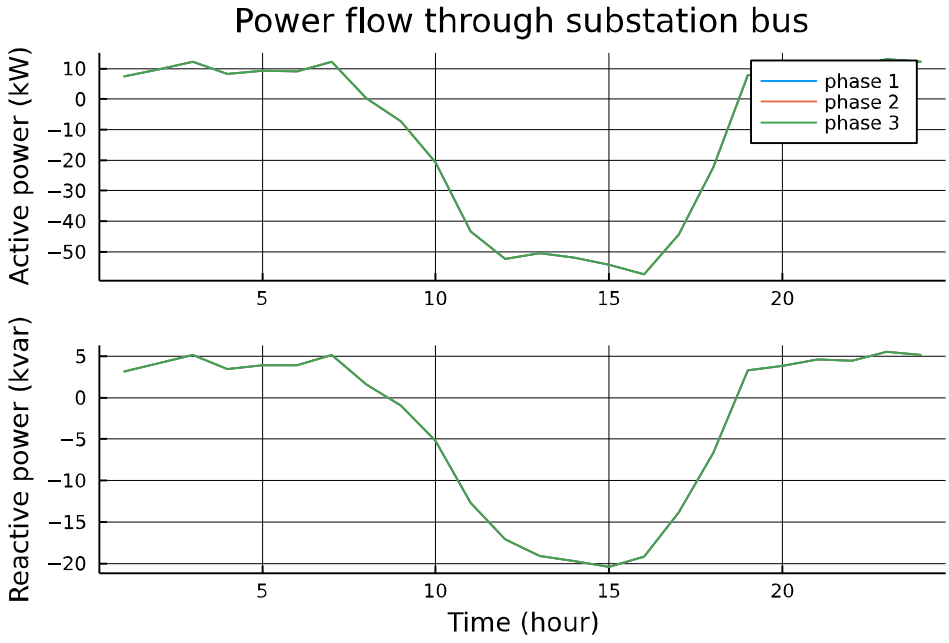


Figure 35 Power flow through the substation, in the presence of PV, in network A

Due to the injection of power, the voltage rises slightly above the reference voltage of 250 V throughout the network. It is noted that in this case, the voltage rise is very limited and despite the high baseline voltage, no over-voltage occurs in this simulation.

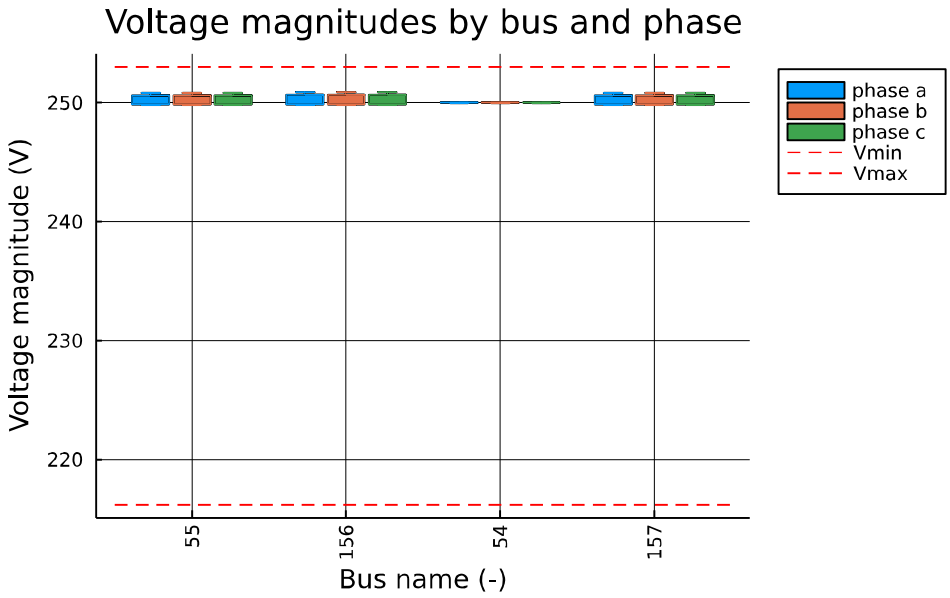


Figure 36 Voltage magnitude boxplots per bus, summarising voltage dynamics in the presence of PV in network A

6.1.2 Network G

Network G has limited unbalance in the power flow through the substation. Nevertheless, unbalance in the voltage magnitudes is more easily observed. As the network operates at 250 V nominally, there is little margin to avoid overvoltage, and observe the overvoltage during the periods of reverse flow. Note that relative to the base case, the lowest voltages are now lower as well, due to mutual induction between the phases.

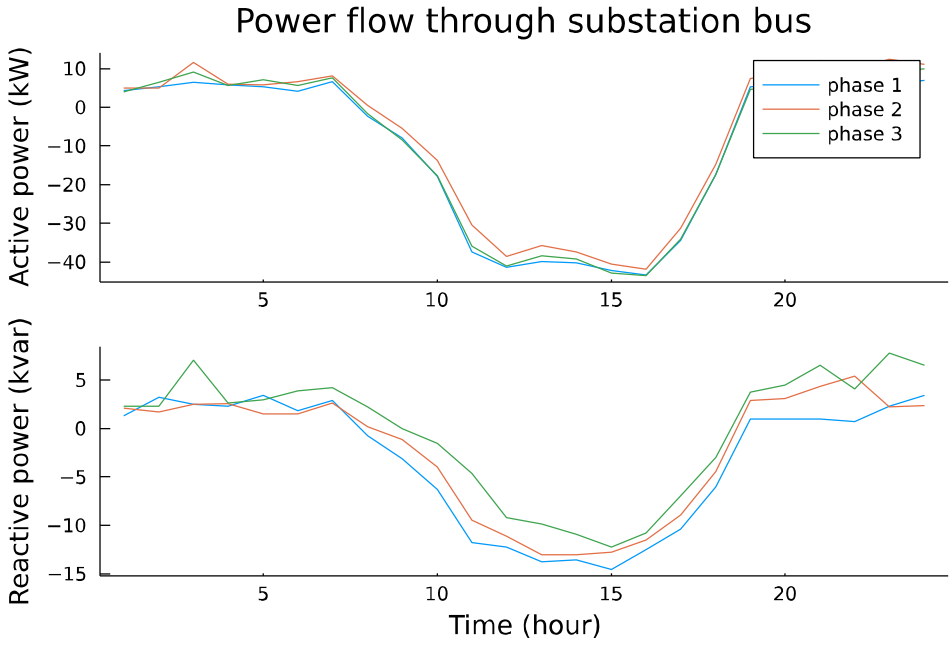


Figure 37 Power flow through the substation, in the presence of PV, in network G

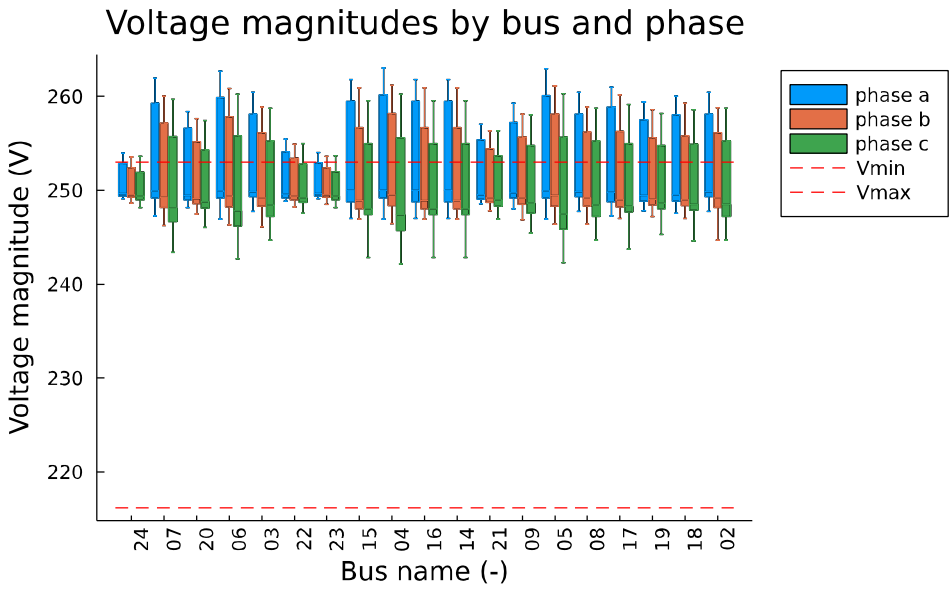


Figure 38 Voltage magnitude boxplots per bus, summarising voltage dynamics in the presence of PV in network G

6.1.3 Network J

Nework J also has significant reverse flow during the day. We note significant voltage rise, again due to a nominal operation of 250 V.

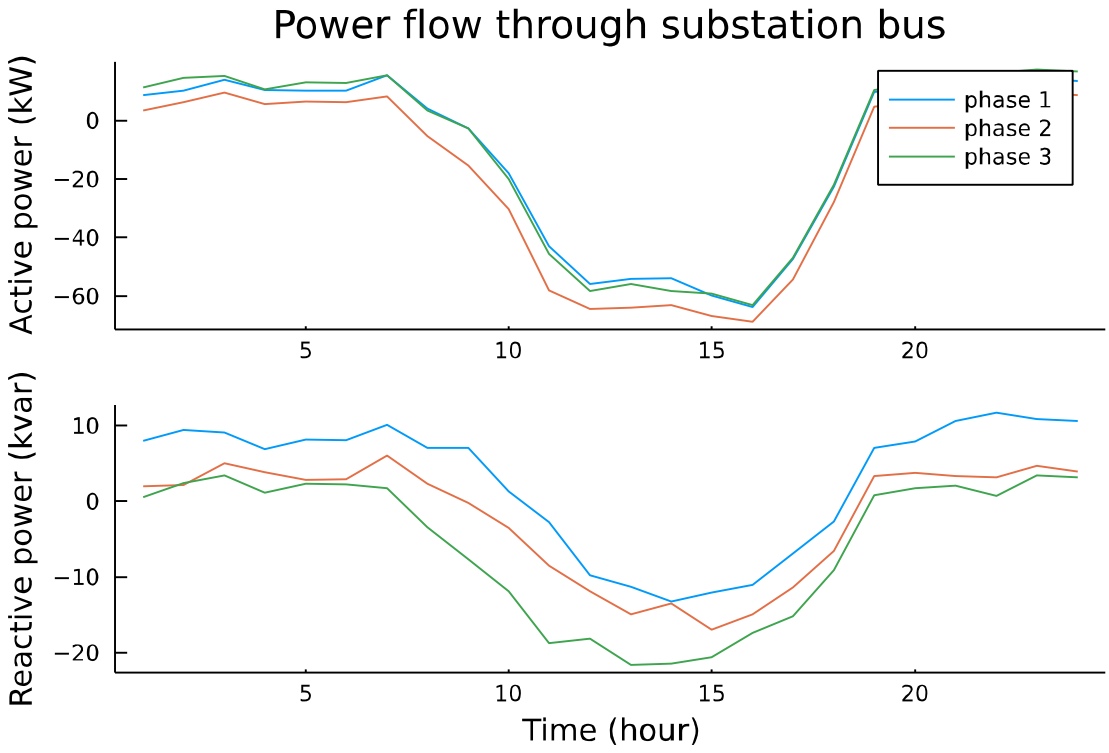


Figure 39 Power flow through the substation, in the presence of PV, in network J

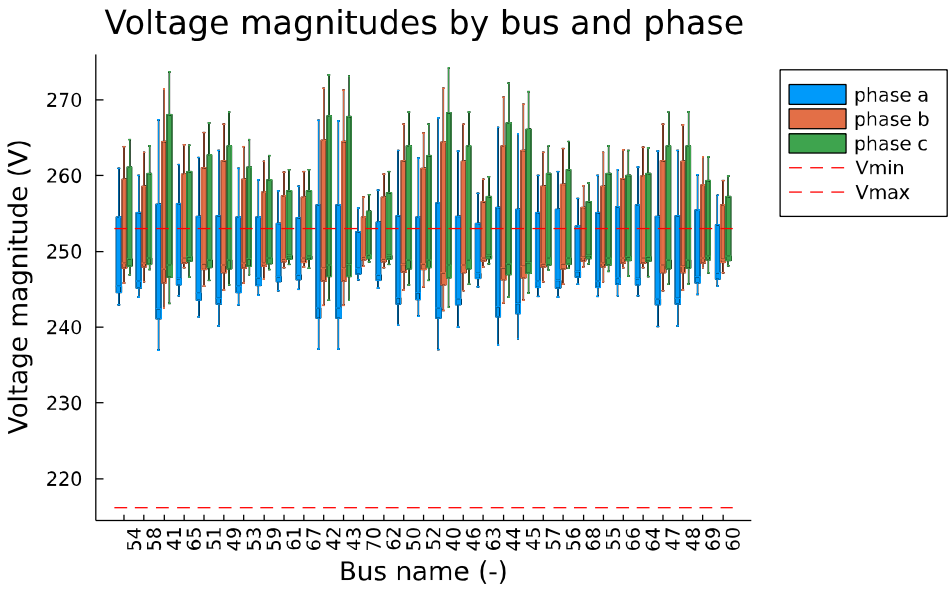


Figure 40 Voltage magnitude boxplots per bus, summarising voltage dynamics in the presence of PV in network J

6.1.4 Network L

Network L has few voltage problems, despite reverse flow..

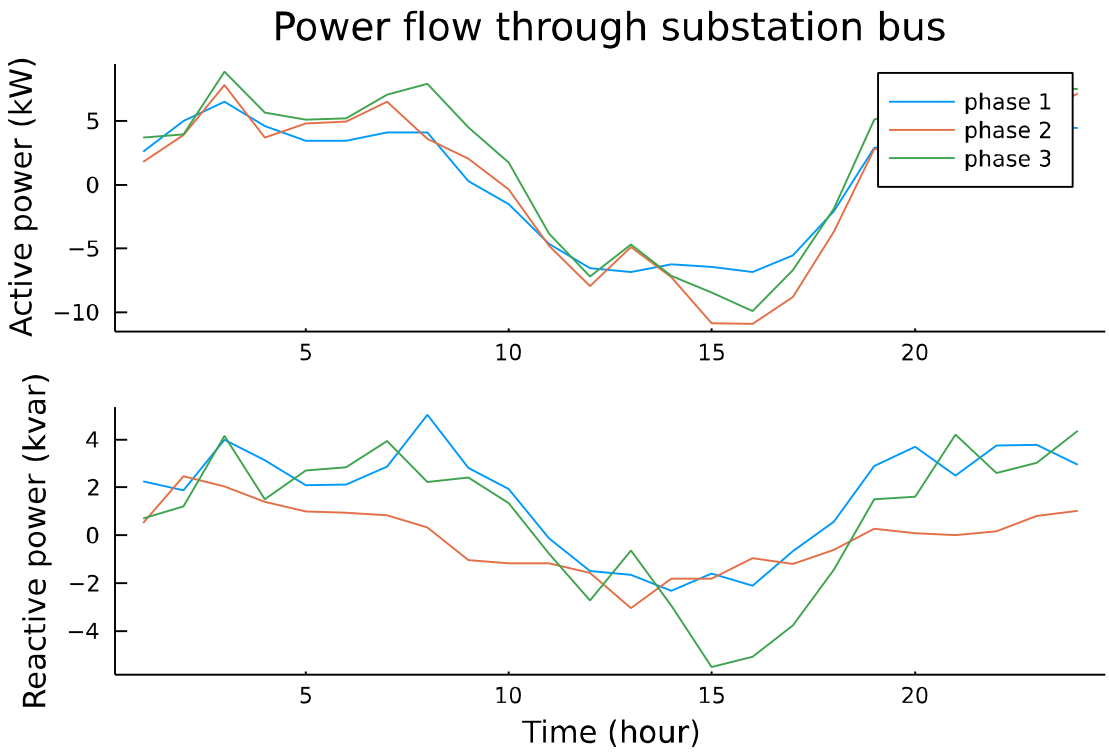


Figure 41 Power flow through the substation, in the presence of PV, in network L

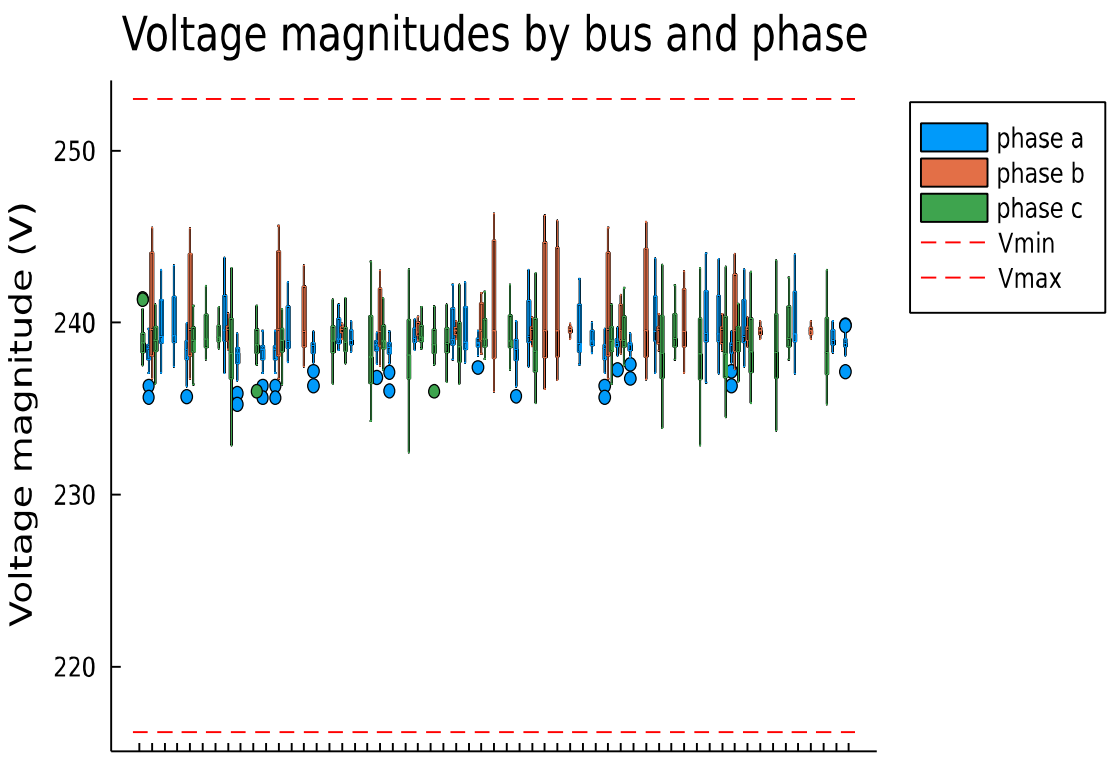


Figure 42 Voltage magnitude boxplots per bus, summarising voltage dynamics in the presence of PV in network L

6.1.5 Network M

Again despite significant reverse flow, there are few voltage problems in this network, due to the nominal voltage setpoint being close to 240 V.

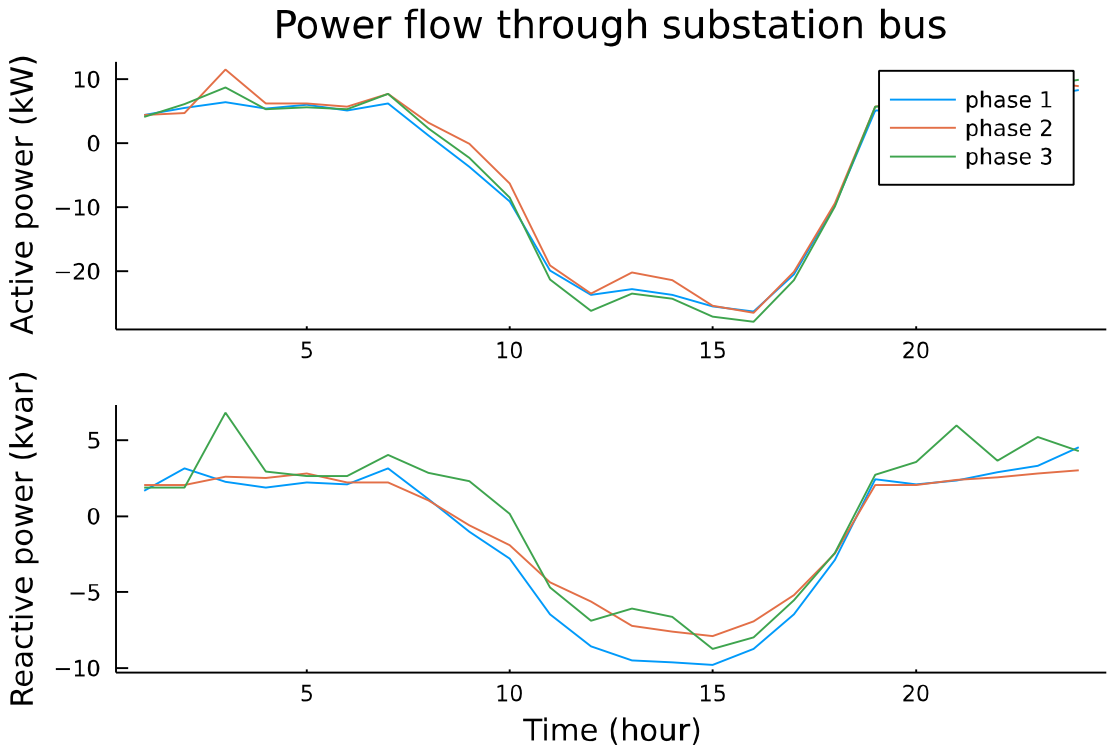


Figure 43 Power flow through the substation, in the presence of PV, in network M

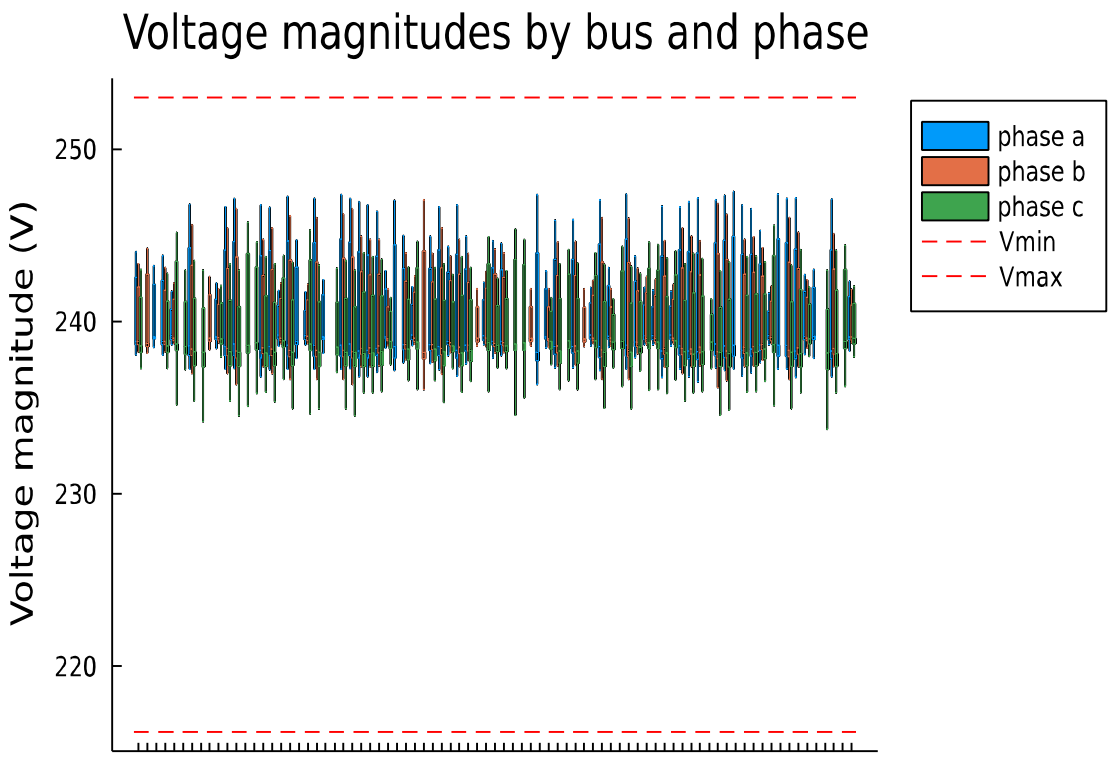


Figure 44 Voltage magnitude boxplots per bus, summarising voltage dynamics in the presence of PV in network M

6.1.6 Network Q

Overvoltage is largely avoided, however, there is little margin left to increase the uptake of PV. In the presence of phase unbalance and mutual impedance, we see that the lowest voltages become lower.

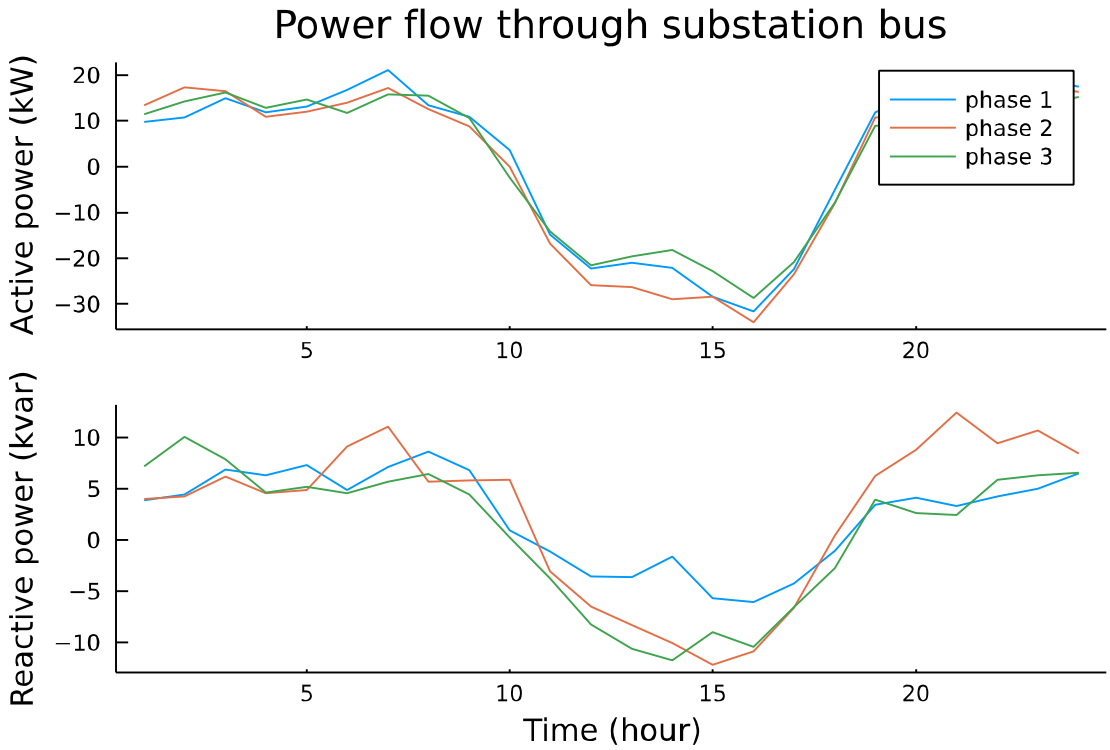


Figure 45 Power flow through the substation, in the presence of PV, in network Q

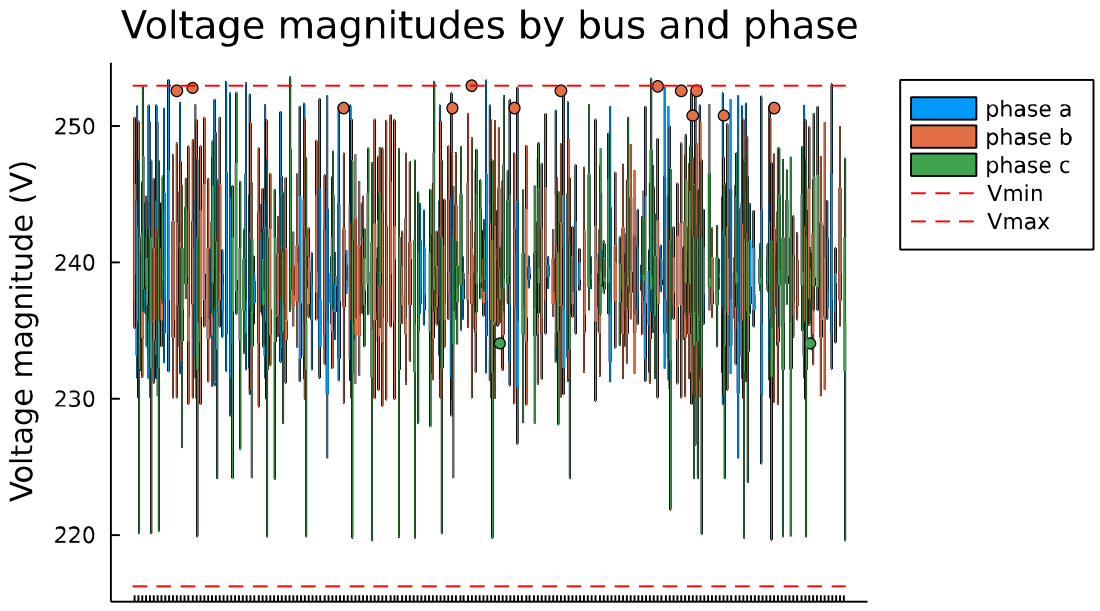


Figure 46 Voltage magnitude boxplots per bus, summarising voltage dynamics in the presence of PV in network Q

6.1.7 Network R

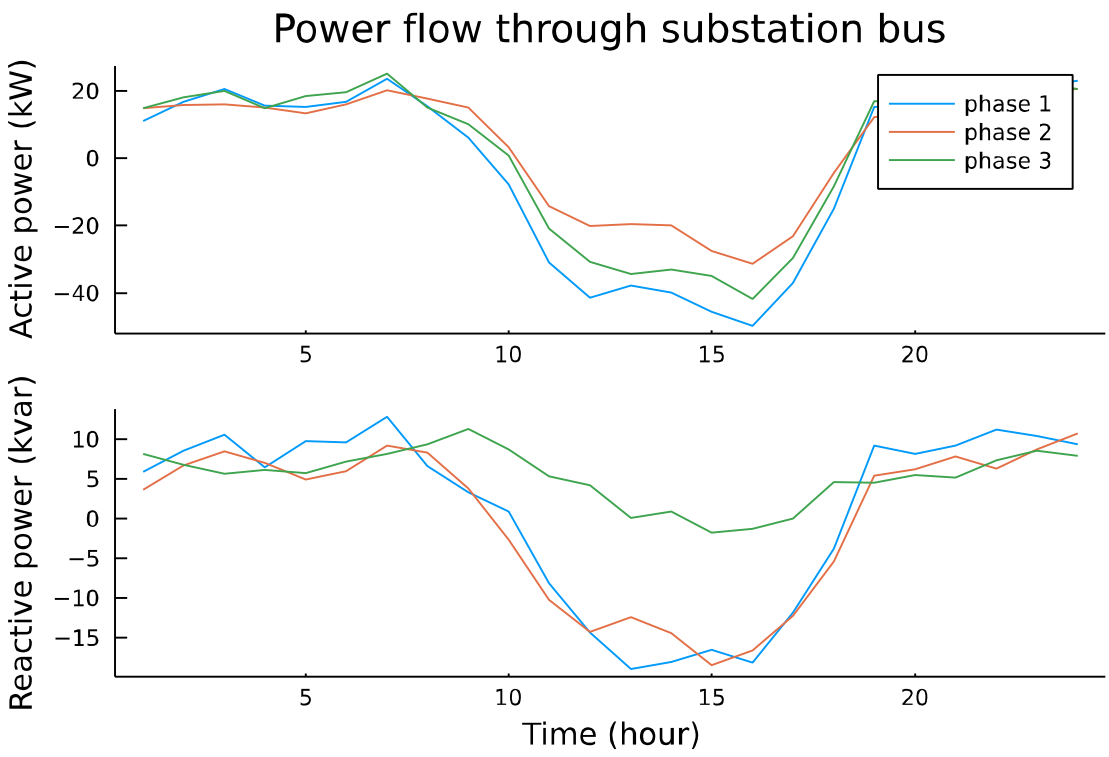
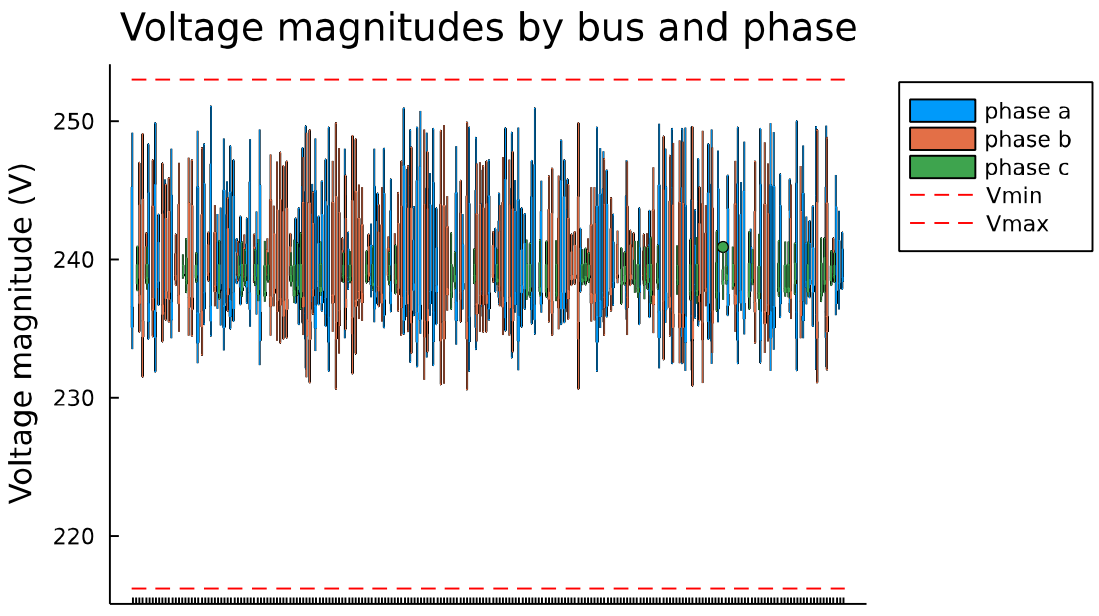


Figure Voltage magnitude boxplots per bus, summarising voltage dynamics in the presence of PV in network R



6.1.8 Network T

In this network, all the loads are connected to phase b on the LV side. Through the delta-wye transformer feeding the network, that gets translated into nonzero flows in phase in the 1st and 2nd phases on the primary. Despite reverse flows, the impact on voltage magnitudes is limited.

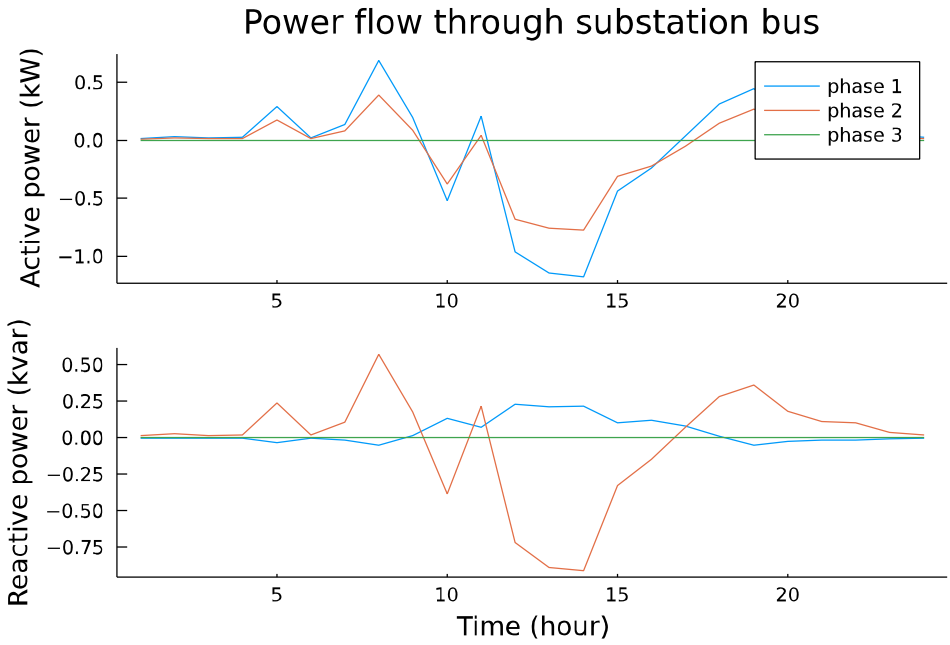


Figure 47 Power flow through the substation, in the presence of PV, in network T

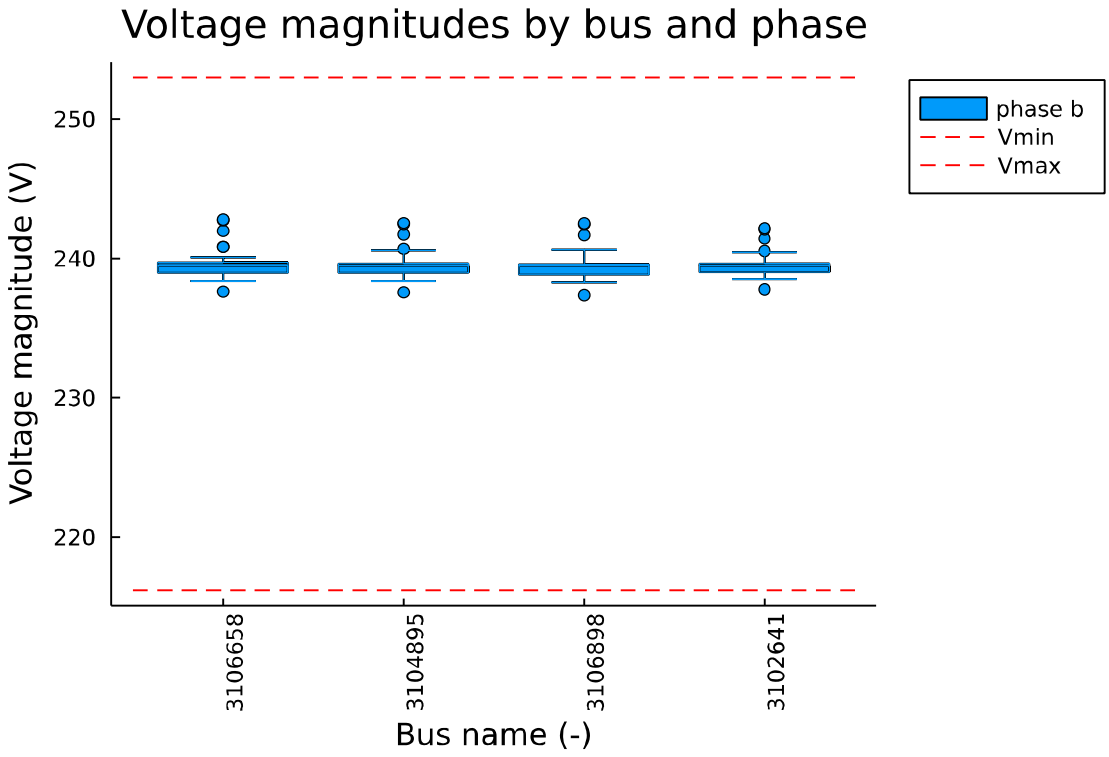


Figure 48 Voltage magnitude boxplots per bus, summarising voltage dynamics in the presence of PV in network T

6.1.9 Network U

We see reverse flow in two phases, but not the third. Voltage deviations from nominal remain limited.

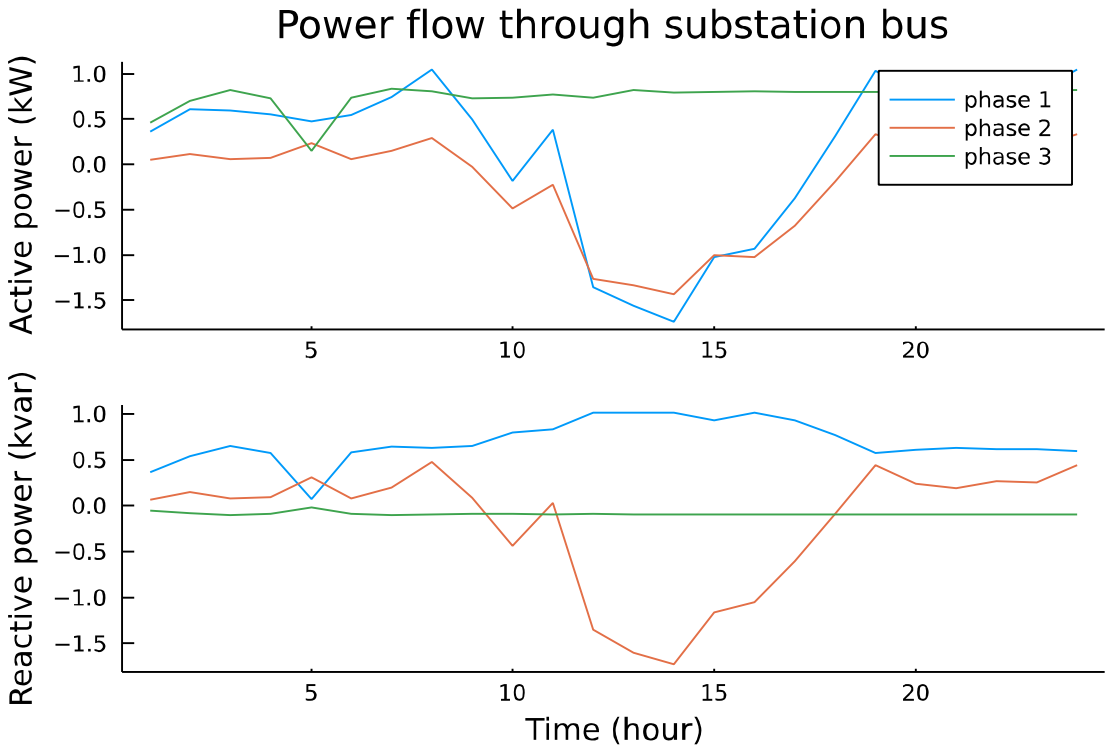


Figure 49 Power flow through the substation, in the presence of PV, in network U

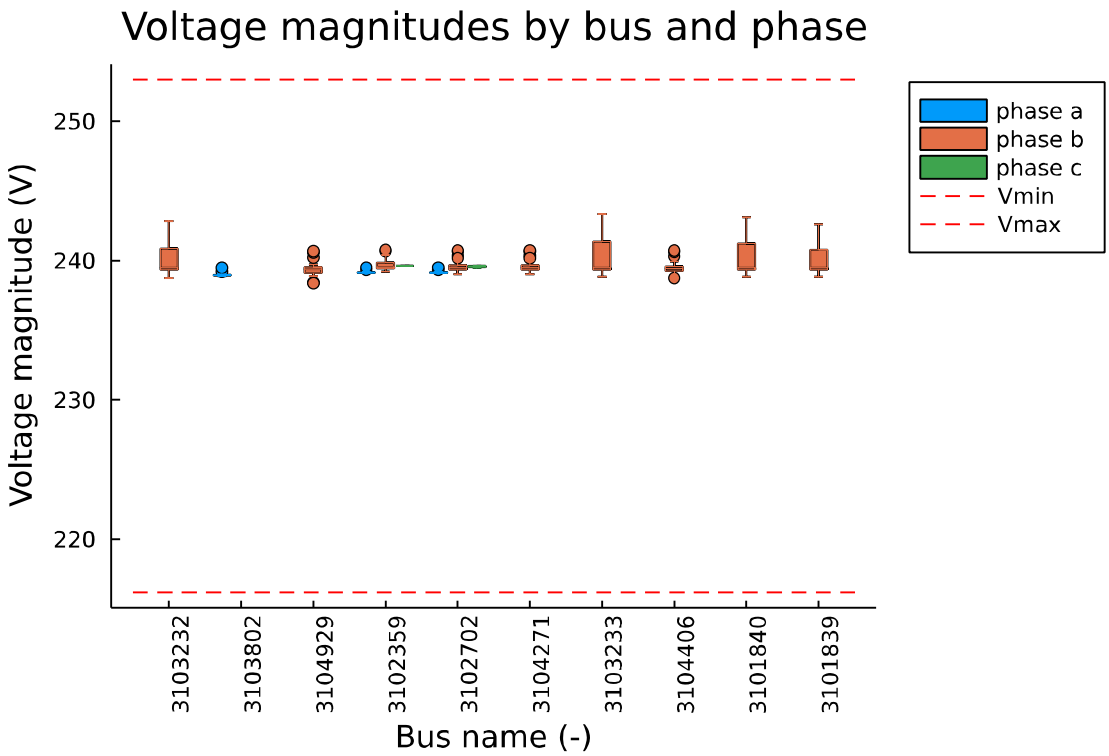


Figure 50 Voltage magnitude boxplots per bus, summarising voltage dynamics in the presence of PV in network U

6.1.10 Network V

Network V has limited reverse flows. The voltage rise is limited to phase a and c, with voltage drop prevalent in phase b.

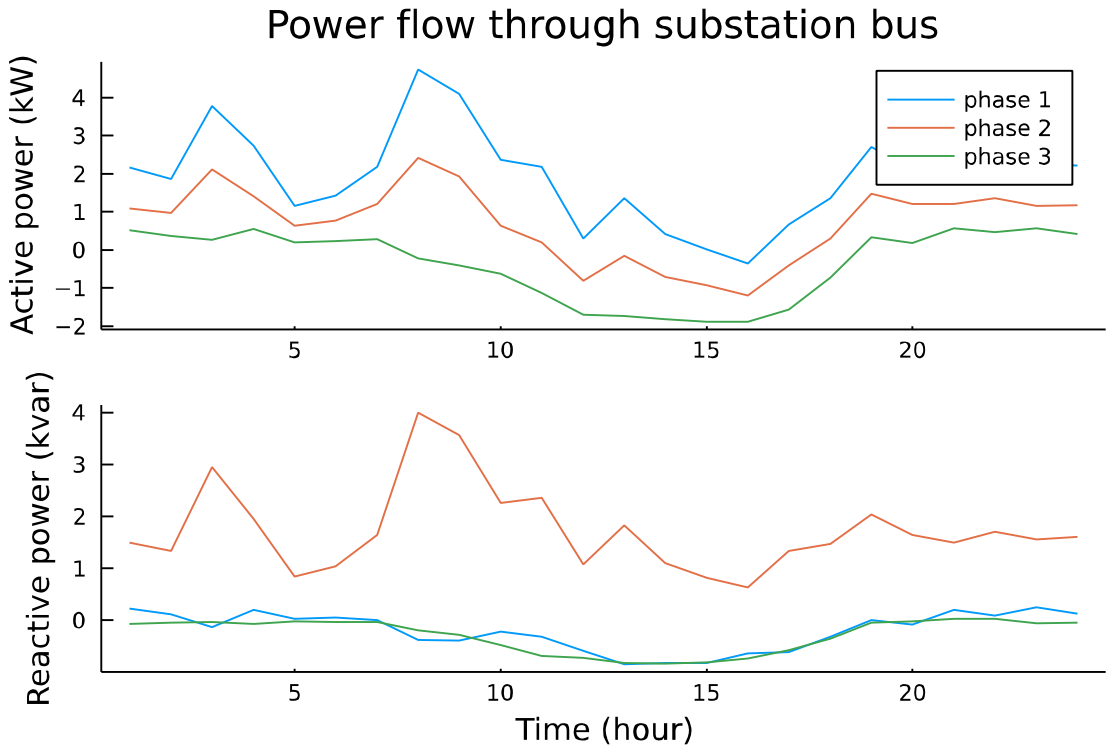


Figure 51 Power flow through the substation, in the presence of PV, in network V

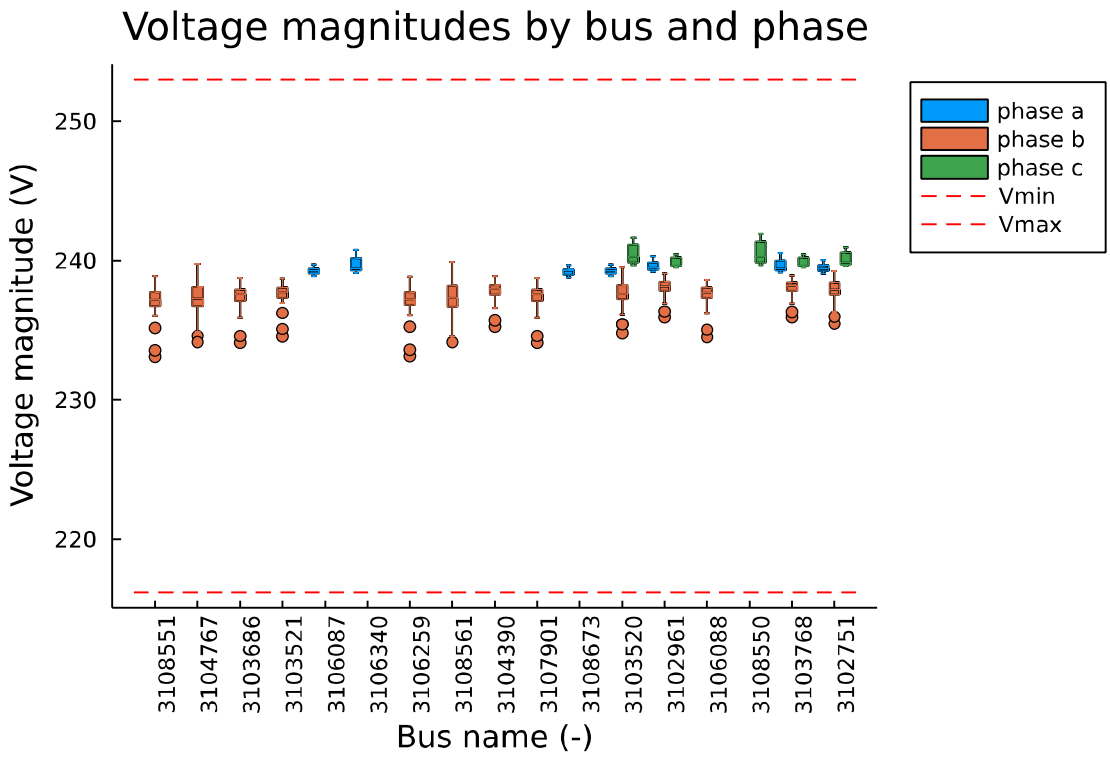


Figure 52 Voltage magnitude boxplots per bus, summarising voltage dynamics in the presence of PV in network V

6.2 Battery storage systems

We add storage systems to buses where there are loads in the base case. We set them up for peak shaving the corresponding load. This will then determine the charging/discharging schedule of the storage units. We use the defaults shown in Figure 53.

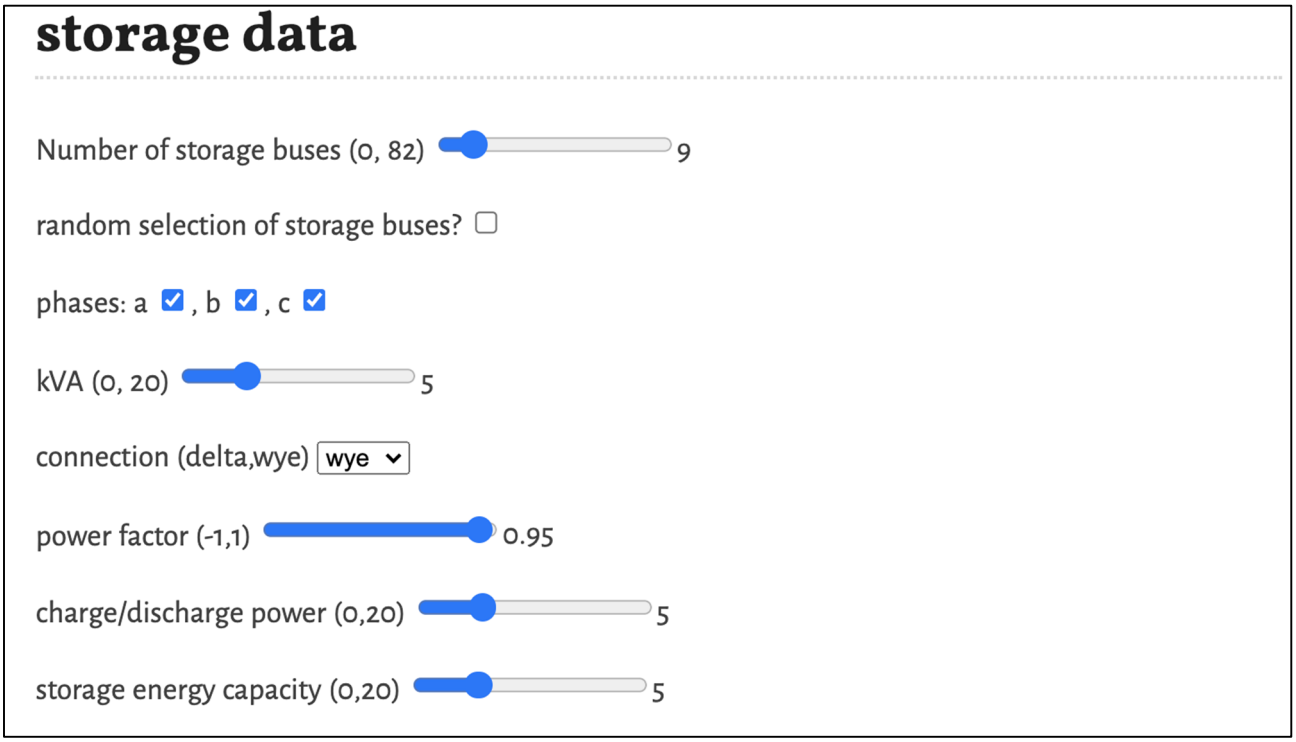


Figure 53 storage system configuration parameters at their default values

6.2.1 Network M

We add 8 storage systems to network 8. Most of them are three phase, but not all (look for the gaps in colour in Figure 54). All the storage systems are dispatched to enable peak shaving w.r.t. an assigned load. Figure 55 illustrates how one of the batteries (the first one in Figure 54) is dispatched to achieve that peak shaving goal. Next, Figure illustrates the changed power flow through the substation.

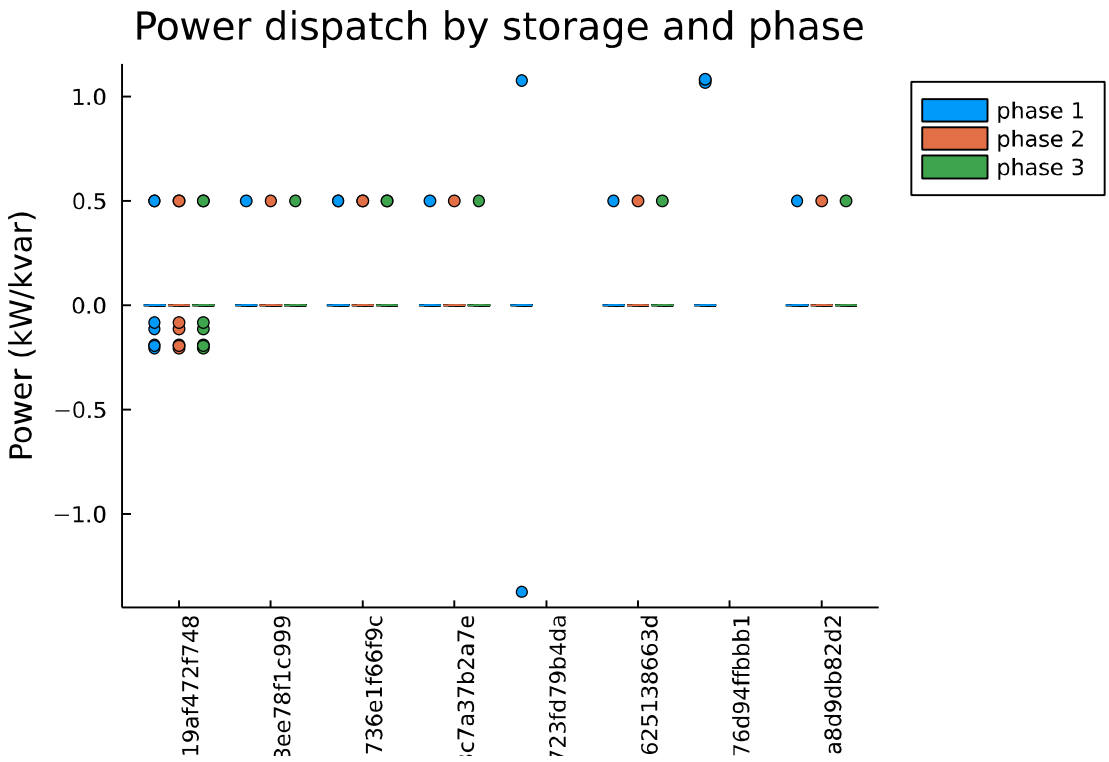


Figure 54 Boxplot of the charge/discharge schedule for all storage systems in network M

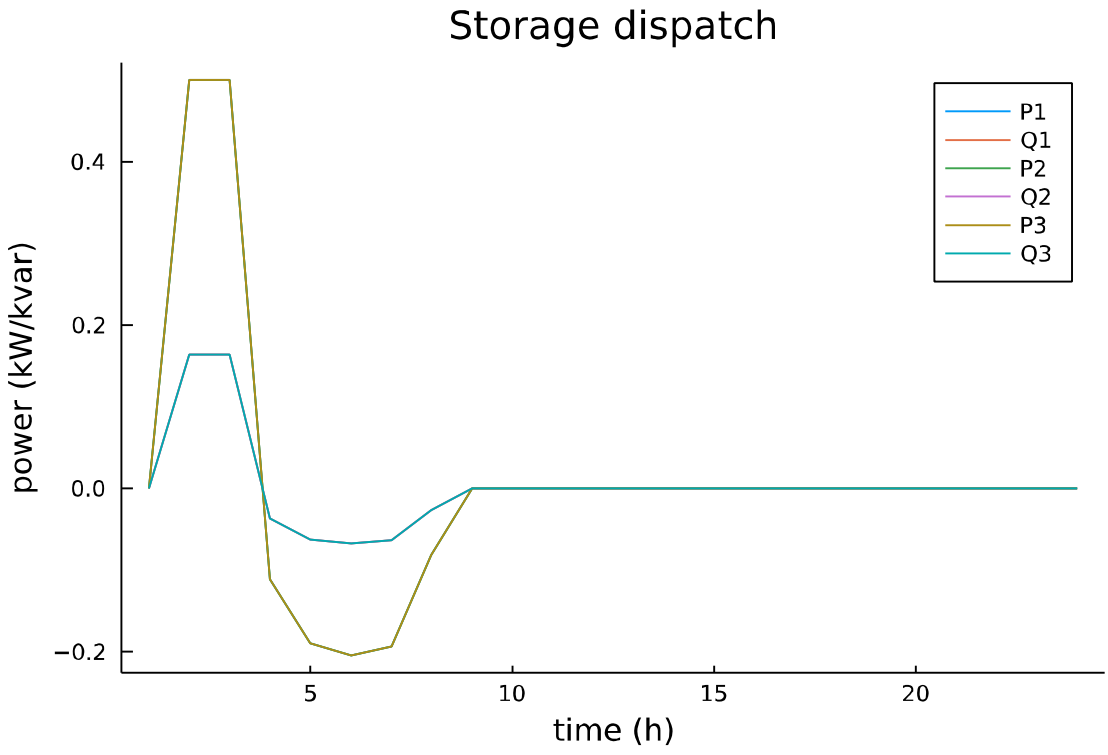


Figure 55 Illustrative charge-discharge schedule for the first storage system in network M

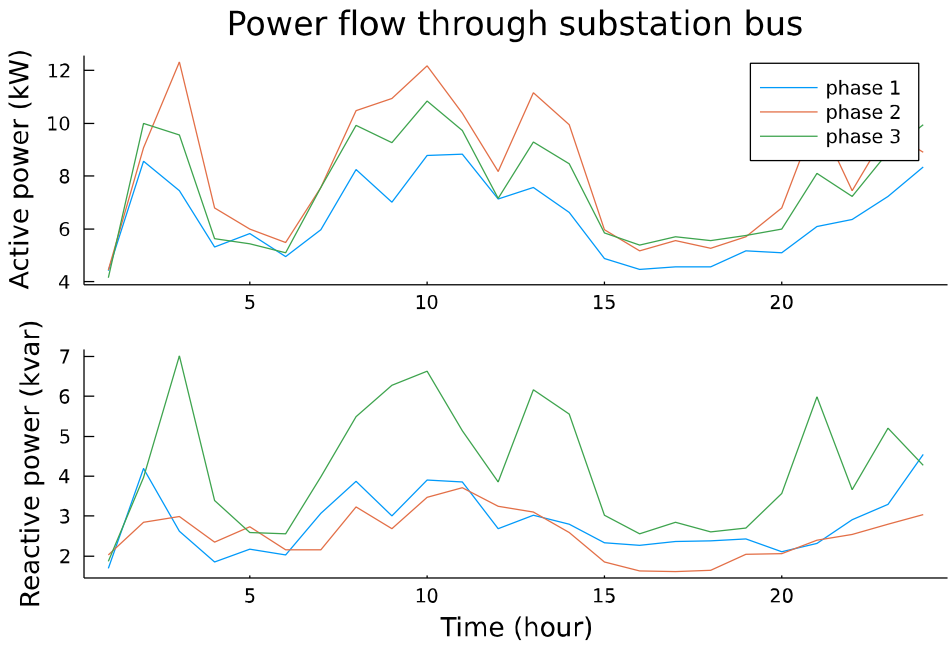


Figure 56 Power flow through the substation, in the presence of storage, in network M

6.2.2 Network O

We now add 13 storage systems to network O. Figure 57 illustrates the charging/discharging behaviour of the individual storage systems in the network. It is noted that all batteries are connected single-phase, to phase a. Figure 58 shows the charge/discharge schedule for the second battery. Figure indicates how the power flow through the transformer changed due to the batteries being dispatched.

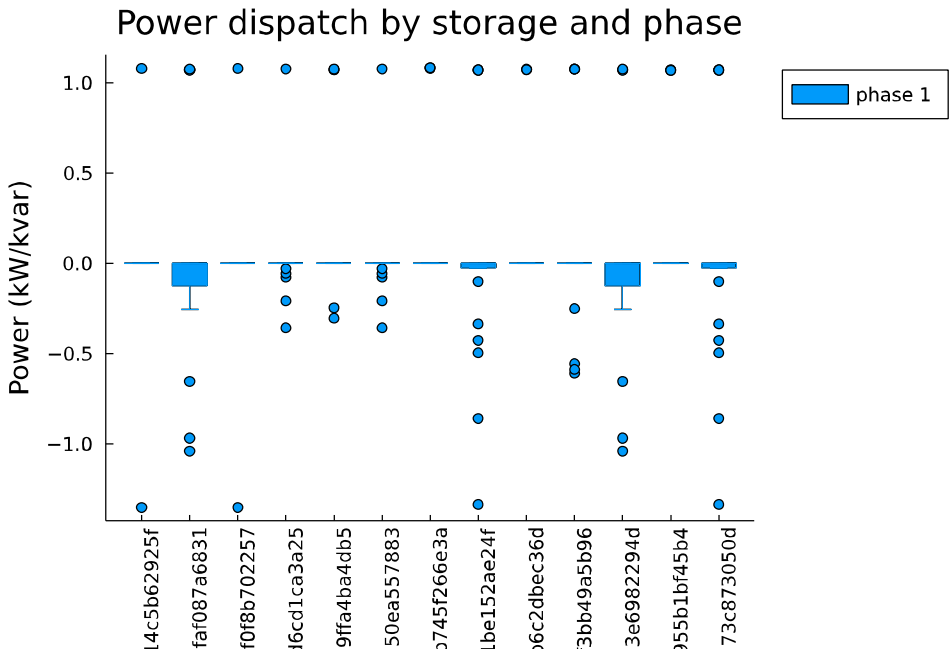


Figure 57 Boxplot of the charge/discharge schedule for all storage systems in network O

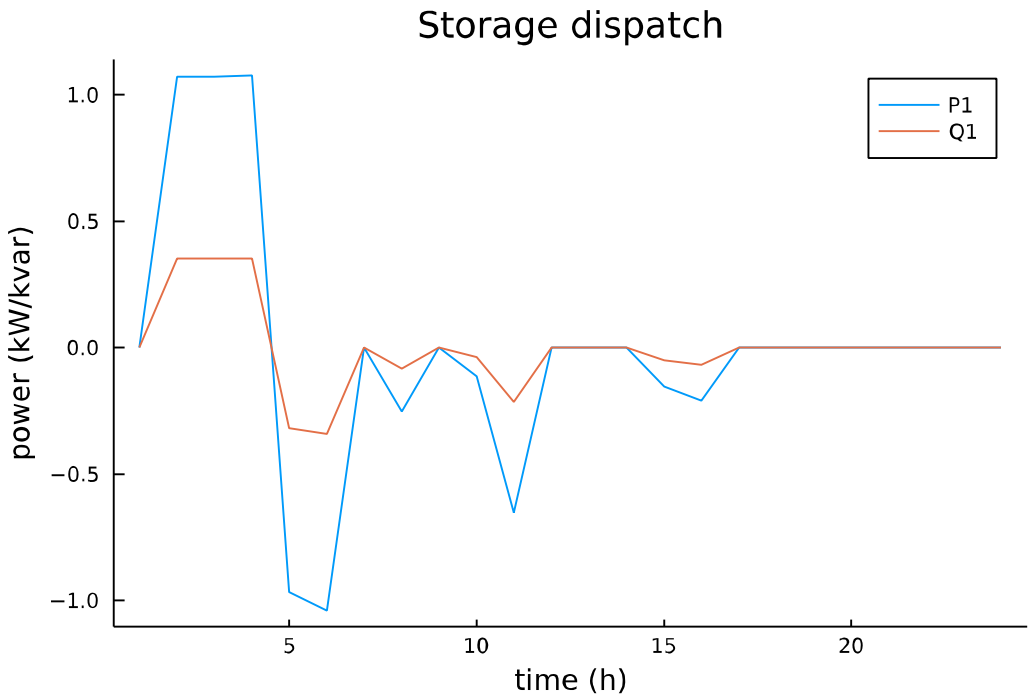


Figure 58 Illustrative charge-discharge schedule for the first storage system in network O

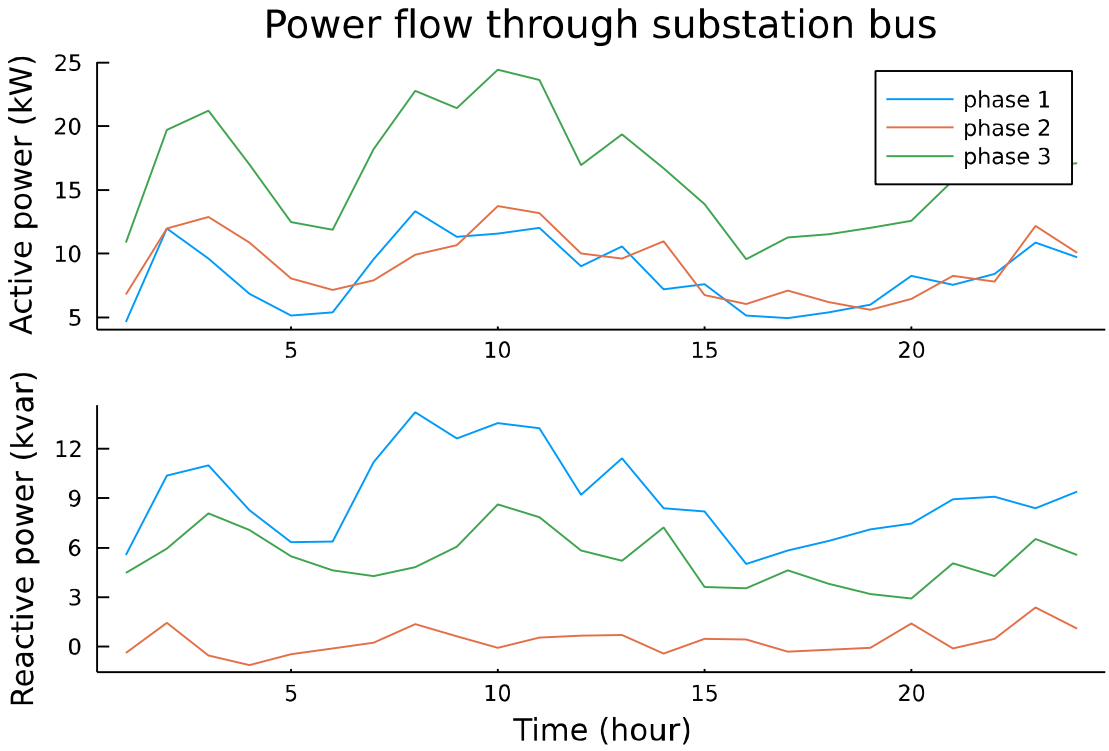


Figure 59 Power flow through the substation, in the presence of storage, in network O

6.3 Demand response

We consider delivering demand response through changing transformer taps and triggering conservation voltage reduction (CVR). Therefore, we change the load model to a voltage-dependent one, in contrast to constant power used in the previous examples, and then experiment with changing the transformer tap (voltage reference set point). This will result in varying degrees of reduced power consumption by loads at lower voltage. We note that trials to enable demand response through CVR have taken place in Australia, see[22].

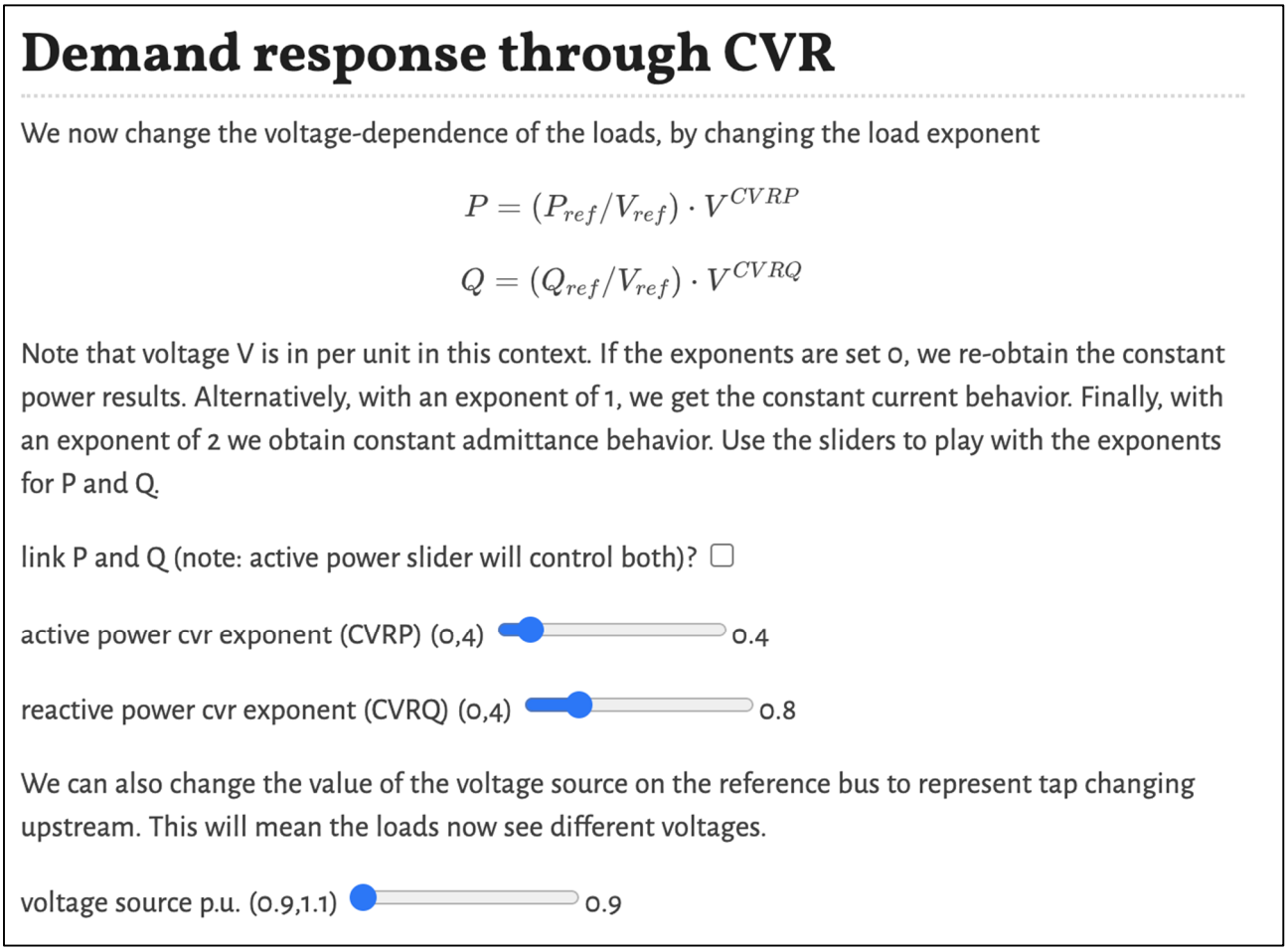


Figure 60 Demand response default settings

[22] *United Energy Demand Response Project Performance Report, https://www.unitedenergy.com.au/wp-content/uploads/2020/01/Project-Performance-Report-Milestone-6.pdf*

6.3.1 Network A

The distance between the blue and green lines in Figure indicates how power consumption in the network would behave under a chance of transformer setpoint from 1 to 0.9 pu, given the chosen load voltage sensitivity factors. Similarly, the corresponding change in reactive power is indicated by the red and purple lines.

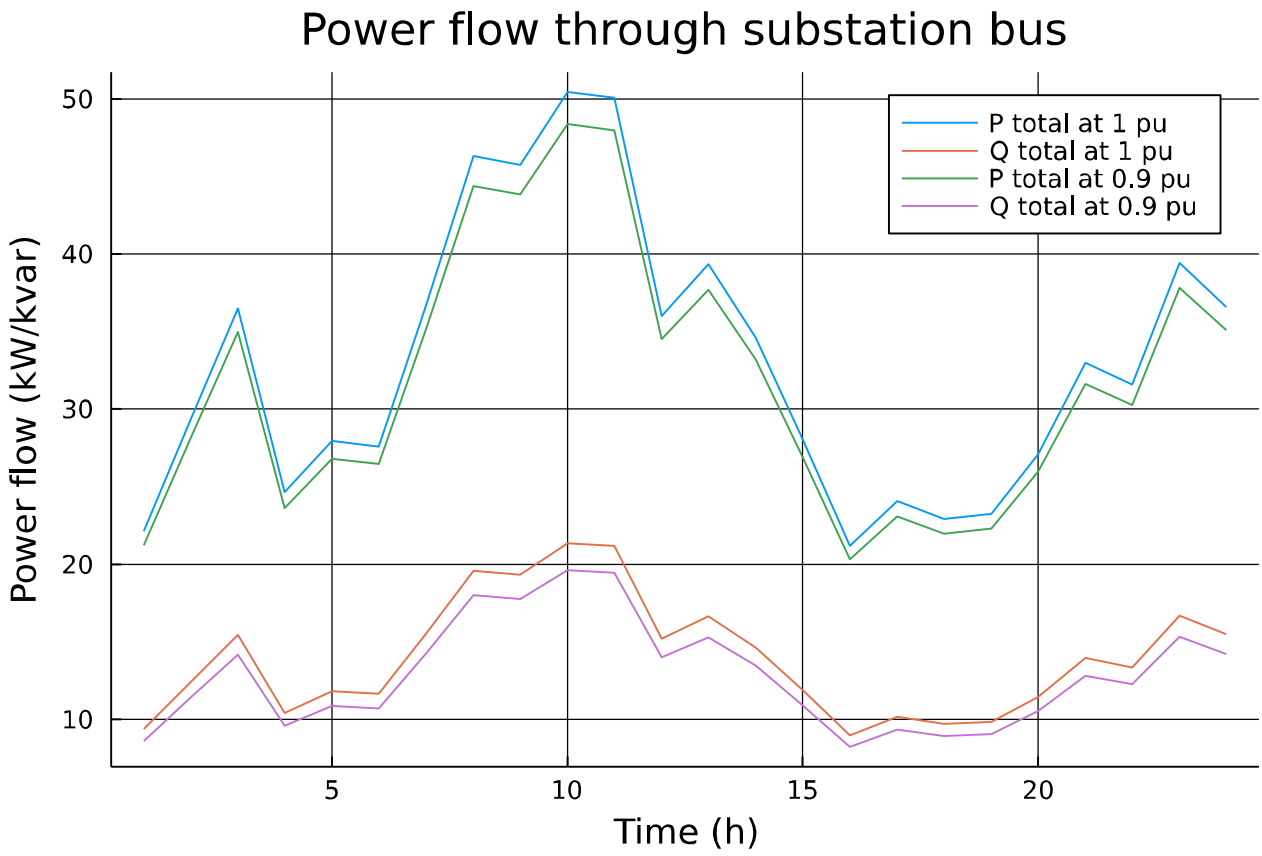


Figure 61 Power flow through the substation with and without demand response in network A

6.3.2 Network B

We repeat the demand response analysis for network B in Figure 62.

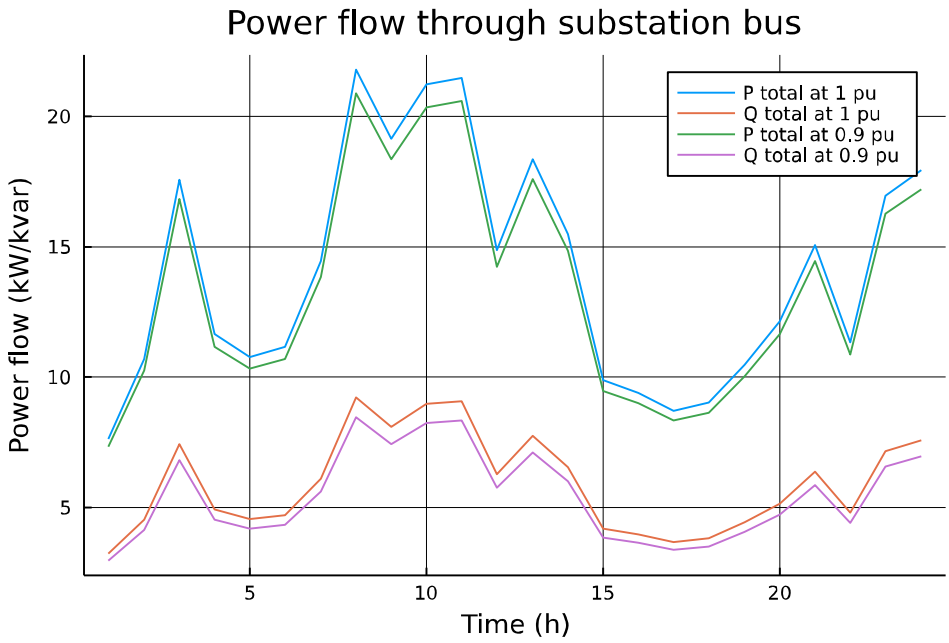


Figure 62 Power flow through the substation with and without demand response in network B

6.3.3 Network G

As a final illustration, we show the effectiveness of CVR to enable demand response in Figure 63.

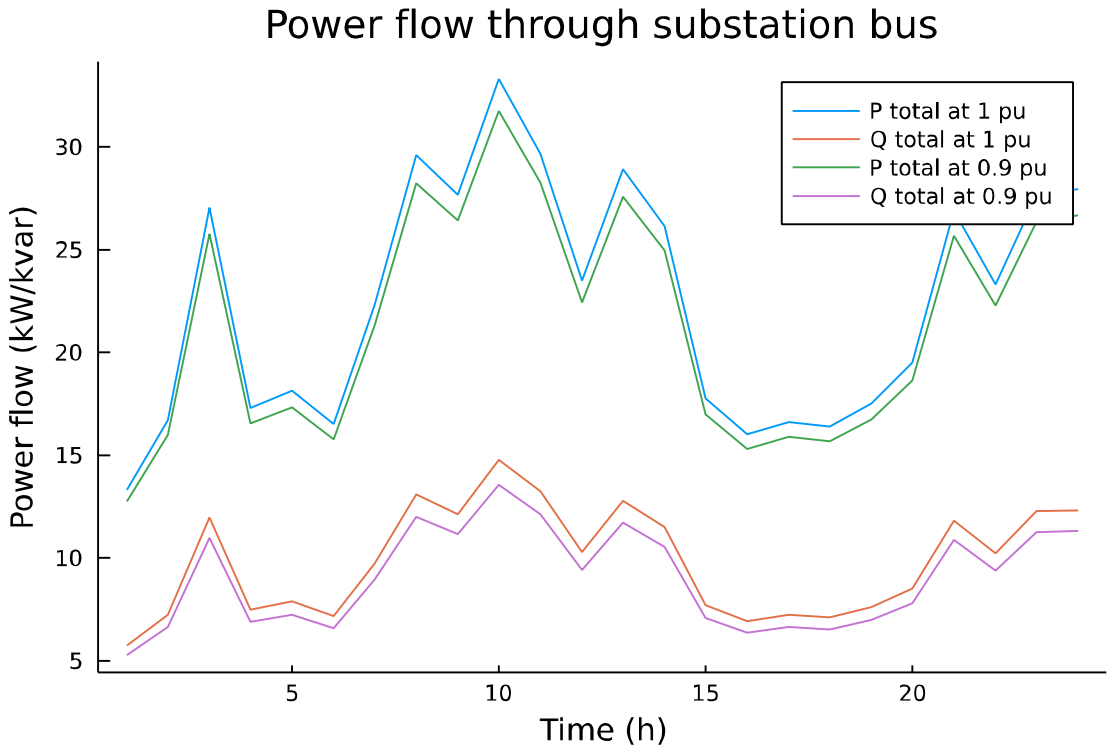


Figure 63 Power flow through the substation with and without demand response in network G

7 Lessons Learnt

This chapter discusses lessons learnt from the project and is split into five categories: re-envisioning the project, use of the taxonomy, data, modelling process, and finally, data sharing and privacy.

Before we provide lessons learnt specific to the LVFT, we re-iterate the 10 lessons learnt from the 'Solar Enablement Initiative Lessons Learnt' report (Krause, 2019). We cannot overstate how much the lessons learnt there were also applicable to this project, namely that:

1) *there remains a need for teams tailored to project deliverables*
2) *software integration is cumbersome and time-consuming*
3) *data quality improvements need to flow back to the original data source*
4) *network data can be inconsistent*
5) *mutable network parameters (e.g. states of switches) need to be tracked*
6) *better middle-ware for tracking network data across different information system siloes (e.g. SAP, GIS, PowerFactory /Sincal) is needed*
7) *data sharing can be cumbersome due to security concerns*
8) *data sharing can be difficult due to privacy concerns*[23]
9) *regular delivery allows for more iteration moments and therefore maximises impact*
10) *DNSP engagements are time-consuming and are hard to track given the combination of large datasets and large number of partners.*

7.1 Re-envisioning the project

To a certain extent the project's more ambitious objective, to develop a national taxonomy representative of **real-world distribution networks**, was a failure – instead we developed a taxonomy of **distribution network data**, based on the incomplete data that the DNSPs we collaborated with have assembled it so-far. That suggests there may be clusters where data artefacts led to the classification, rather than features from the real-world networks. We attempted to mitigate the risks here for the representative networks but could not achieve completely artefact-free data sets for clustering. The relatively simple, static, metrics used for clustering (due to them being required for the majority of networks) may be less relevant than power flow simulation results (required on all network models) to determine which networks are most relevantly similar. Future efforts should focus on ensuring that all network models can be simulated (after conversion to other formats) and are comparable (using appropriate time series data), so that power flow features can be used for clustering as well as the other features.

An important thing to note is that we produced a low-voltage network taxonomy, rather than a feeder taxonomy.

[23] Although not in the original source, legal constraints: contracts and governing legislation, can also hinder data sharing

We developed the taxonomy using data science principles, where we first ingest all the data, and then let the data tell us what the clusters and representatives are. Alternatively, the taxonomy could have been approached prioritising domain knowledge from power engineering, as follows. Common network architecture designs would first be identified, e.g. city network with apartment buildings, suburban network with individual dwellings, etc. In a second step then we find representative examples for each of those categories. Using automation to consistently identify such network architecture designs across different DNSPs is not straightforward, and would require significant data cleaning and tagging algorithm research.

Due to the observed data limitations, two fundamental questions arose during the project: What is the purpose of a taxonomy and is it worth the effort required to develop them? We address these questions by re-thinking how to increase the scope and applicability of the taxonomy. An ideal outcome would be that the taxonomy facilitates the extrapolation of the performance and costs of new technologies from category representatives to wide-scale deployment. To do so with confidence, we would need:

- further cataloguing of biases, and sensitivity analyses to understand their impact on the resulting clusters
- a better understanding of variance of the relevant clustering features in the real world, as would be achieved by careful statistical analysis on the network data. A key challenge is that we don't yet understand the frequencies of occurrence of wide range of broader network architectures and features. A number of real-world highly relevant features, that were associated with high quality data for some network, were dropped due to them being missing for other power networks within the dataset.
- additional network data that covers a broader scope of *national* coverage.

Being able to confidently extrapolate from an analysis of cluster representatives requires knowledge of the proportion of the target population that is represented by each cluster. Furthermore, one should then perform sensitivity analysis of being able to extrapolate previously unseen features. It is to be expected, for example, that a cluster based categorisation that is usefully predictive for one particular given performance indicator (eg susceptibility to thermal overload) may not be a useful categorisation for another (eg poor power factor).

The previously identified, more general, recommendations would still need to be followed: more quality control on data processing pipeline, use power flow metrics as part of clustering, make progress on algorithms for harmonising and cleaning network data sets. As networks change due to the integration of PV, batteries and other new technologies, re-doing the statistical analysis would also be prudent.

7.2 Use of the taxonomy

We note that a number of questions can be raised with respect to the use and applicability of the taxonomy within DNSPs:

- How do we extrapolate from this taxonomy to identify the remaining hosting capacity in Australian distribution grids?
- How should DNSPs use the taxonomy as part of the network planning process?

- How can novel energy technologies be validated in the taxonomy?
- How does the taxonomy help us design appropriate trials of innovative technologies?

We do not recommend using the representative networks for extrapolating hosting capacity across networks, as we have no idea how representative the occurrence of congestions is. Their usefulness as a testbed serves to inform on the different options for technology integration, doing simulation trials of interventions to improve performance. Furthermore, the representative networks can help design and develop rules-of-thumb, e.g. for sizing of the network conductors in networks with a high amount of electric vehicles. However, in the context of network planning, extrapolations for deployment costs of novel solutions are ill-advised.

The validation of new technologies in simulation is hard. The appropriate methodology depends on the features of the proposed technology. The application of the developed steady state unbalanced power flow simulation can provide a useful initial approach to validate the efficacy of new technologies, but it may not be enough. Power engineers will often need to run other kinds of analyses as well to gain confidence in a novel technology or to design trials, such as:

- short circuit simulation,
- harmonic power flow,
- electromechanical transient simulation,
- electromagnetic transient simulation,
- network-constrained optimal dispatch, e.g. for coordinating the curtailment of PV, assigning export limits.

It is noted that the data release has not been designed to enable those other simulation problems to be solved without modification or extension.

7.3 Network Data Curation

We note that a few DNSPs had historically already produced their own taxonomies, based on real-world and/or expert insights. Taxonomies or archetypes are an interesting option for DNSPs that currently do not have any low-voltage network models. Nevertheless, they are only a stop-gap solution. As we do not know how representative the power flows and congestions are for the real world, this taxonomy should be used as a testbed only, not as a basis for extrapolating the hosting capacity of Australian distribution networks. The purpose of a testbed is to inform and facilitate early exploration in simulation related to the integration options of new energy technologies. The clustering process should have included power flow results if we wanted the outcomes to be representative for network congestion as well. But that would have required significantly greater data cleaning, and possibly working more intensively with DNSPs to make data consistent across them.

Cleaning network data at this scale, across different DNSPs and different proprietary formats is a software challenge that needs to be engineered for, with the appropriate quality assurance processes. Advancing the software stacks in this space is needed to enable data quality suitable for creating digital twins. We note that cleaning data on networks that serves as an input to simulation engines is particularly challenging, as errors or misconceptions in the data transformation and processing may result in a model that is unsatisfactory for simulation.

Furthermore, those bugs are hard to find and solve without domain expertise. Interactive network visualisations can be helpful for cleaning and troubleshooting. Keeping track of the data from the source to its final processed form was also surprisingly challenging.

Proprietary data formats, such as the PowerFactory '.pfd', are a significant barrier, particularly when they require expensive licenses. Furthermore, there are no commercial-grade tools to convert between different formats, or to open them up. OpenDSS supports the description and simulation of almost any practical power network configuration, with the maximal amount of detail for steady-state simulation. However, this flexibility also makes the learning curve relatively steep. Furthermore, power engineering tools rarely include functionality to debug network data. There are opportunities to define more structured data models, that are easier to conceptualise quickly. For instance, this could be a data model that fits only 4-wire LV networks or one that only fits MV SWER systems. When such structure is known, it should be easier to automate data quality control checks and feedback.

A custom algorithm was added to the DiTTo reader code to split the original MV networks into multiple individual LV networks. In hindsight, it would have been preferable to add this splitting functionality as a separate function rather than making structural changes to the DiTTo parser, which makes it difficult to contribute these changes back to this open-source project.

Another challenge is that network data should be maintained so that it is consistent with the network in the field. If lines are replaced, added, or re-configured, the augmented design should be updated in multiple software systems, i.e. GIS, ADMS and power engineering simulation platforms such as PSS/Sincal or PowerFactory. Furthermore, over time, the best practices for developing and maintaining network (simulation) data may improve, so the software systems would ideally allow for this in a user-friendly manner, while keeping a historic record of the changes. We note that there is a significant diversity in legacy and best practice approaches to network modelling. The synchronisation of network changes occurring in the field and data updates in software tools implies a greater coordination of network business processes, particularly between line crew and office staff. Anecdotal evidence suggests that process changes and improvements in work flows will prove as important and challenging as – if not moreso than – the development and application of better software solutions.

There are numerous known biases in the final set of networks that we could not compensate for – for example 1) some LV networks don't have service lines 2) some LV networks only model 3-phase loads 3) some DNSPs have only digitised a few LV networks. There are also biases we are unaware of. The number of networks (sample size) from each DNSP varies widely and is not representative of their proportion. Furthermore, they are concentrated in Australia's south-east. We have to assume a strong survivorship bias for the networks where the data was partial: why is the data there or not there? This is likely explained by a process of prioritisation, which suggests, but does not definitely indicate that congested networks may be over-represented. Furthermore, the network modelling practices of DNSPs are different, even when they use the same software. There is a risk that a few of the clusters are selected accidentally for data artefacts, yet these artefacts do not represent the real world, and therefore the cluster or archetype is not representative of the underlying features in the real world.

Power flow engines tend to have edge cases that result in failure to solve the simulation. For instance, lines with zero impedance (alternatively zero length) are a problem. Nevertheless, such

parameters exist in the data provided to us by DNSPs. This meant we had to write algorithms to filter out these problematic components. Similarly, keeping track of the open/closed status of switches, breakers, and sectionalisers is a challenge. The network data provided largely indicated the design structure of the underlying networks, rather than the real-life configuration of the networks in operation. Therefore, we also had to use heuristics to find likely radial network configurations.

A key recommendation to clean up the network data is to develop novel first-principles cable and overhead line libraries, in the 'abcn' instead of sequence component coordinate space. Ideally this is based on the type (overhead open wire, bundled, underground cable, concentric neutral etc.) geometry, material and cross section information, so that Carson's equations can be applied (Kersting & Green, 2011). This information is also the building block for building harmonic power flow models. Furthermore, linking this up with specification sheets in visualisations of the cross sections, permits the future development of impedance data based on finite element simulation. Finite element-based workflows may offer better accuracy in the context of higher harmonics (n>5).

Finally, the load data wasn't harmonised. We would ideally have access to public, realistic, customer load data with high time resolution, with both active and reactive power profiles. Such data currently does not exist for Australia. Most data sources provide only active power and 30-minute resolution, or they are too limited in diversity. High time resolution is needed to assess network losses accurately in the context of phase unbalance and a limited number of customers (Urquhart & Thomson, 2015).

Network data are (mathematical) graph-based data sets. This means there are features that correspond to edge/line properties, e.g. lines and transformers, and features that correspond to vertex/node features, e.g. voltage limits. Processing such data requires understanding of the topological information. Most of today's data scientist and the tools they use are focused on time-series data and may not yet be up to the task of dealing with such graph-based data sets.

7.4 Workflow design

Data verification and cleaning was required at several stages of parsing and transformation, as some incompleteness or errors were easier to detect and correct in processes further downstream than upstream. The workflow became a pipeline of data transformation steps: reading, projecting onto the target data model, splitting/merging, feature extraction. Cleaning between each transformation step was necessary, with different data cleaning processes depending on the data source: the data from different DNSPs would have different types of error. It became necessary to develop the data extraction and transformation process and software concurrently with performing the data extraction and transformation process itself, as testing on a subset of the source data invariably failed to identify all the data idiosyncrasies. This became an iterative development process that required both software and processed data to be version controlled. The data extraction workflow eventually became fully automated, although during earlier testing phases portions of it were often performed manually.

Although it may have saved a small amount of processing time and memory by avoiding parsing of some parts of the MV network which were not necessary for this project, it became necessary to introduce changes to the base DiTTo classes that will overly complicate reintegrating these parsers

with the main DiTTo code-base. It would have been better to implement simple non-LV-splitting parsers, and then add splitting functionality to the resulting single large network as functionally separate processes. This would have been particularly beneficial for the PowerFactory parser, which relies on a proprietary Python interface to a running PowerFactory instance which can be accessed by only a single thread/process at a time. A single PowerFactory file parsed into an in-memory DiTTo network model, on the other hand, could be split by functions that took advantage of multi-processing, and perhaps more importantly, would have been re-usable for arbitrary input file formats.

7.5 Modelling process

This project should have been conceptualised as the first step in a series of projects. For example, first we try to understand the properties and limitations of the existing network data: perform statistical analysis, explore biases and identify limited-scope representatives. Next, we focus on understanding the gap between the existing network data and what we know about real-world networks from the electrical engineering perspective. Finally, we develop data cleaning and transformation tools to enable clustering approaches for the real-world representativeness. To some extent, we only succeeded in identifying clusters based in the features that the DNSPs have already assembled. That doesn't directly imply that these networks are representative of the real world. For instance, we identify a few gaps there:

- Impedance data is likely inaccurate
- Phase connectivity is generally missing or arbitrarily assigned

The key take-away here is that there is a need to develop best practice documents, workflows and tools for managing, maintaining, cleaning and improving network data over time, while keeping it simulate-able.

Simulation of physical systems serves to inform decision making. The more realistic and representative of the real-world, the stronger the decision basis. Therefore, using data of an actual network is generally preferred to using one that looks vaguely similar (e.g. part of the same cluster), all else being equal. Therefore, we believe that the way forward is to increase the scope of simulation of real networks. Given the data challenges identified, that may not be easy. Nevertheless, progress can be made, engineering tools to clean the original data and automate hosting capacity analysis, and researching tools to anonymise network data sets for public release (including up to differentially private).

Finally, we must reflect on the attitudes around power flow simulation model development. It is often said that there is no such thing as a perfect model, and that it applies to power system simulation just as well. While there is a lot of truth in those statements, the importance is exaggerated in the context of the steady-state physics of power networks. The mathematical equations that govern the steady-state physics, and ways to derive network data from first principles (i.e. Carson's equations (Carson, 1926)) have been known for close to 100 years. What has changed is the technology to solve these mathematical models at scale, i.e. simulate the networks.

Historically numerous simplifying approximations were developed to make the equations easier to solve by hand (i.e. by humans), however with the advent of digital computers in the second part of

the 20th century, we have been able to solve systems of equations at larger and larger scales, thereby being able to simulate larger and larger networks. Nevertheless, some of the approximations to simplify calculations have persisted. However due to advances in computation power, such short cuts are less necessarily. We generally know what are modelling best practices (at least in steady state), and we know how to simulate networks at that best-practice level of detail. While computational challenges exist, they are generally manageable, and we can still (even automatically) apply the well-known approximations if required. Therefore, we should think about hierarchies of simulation models, with better and worse representation of the physics – putting different models of the physics at equal footing is often misleading. For instance, if sufficient data is available, an *unbalanced* power flow simulation is inherently more accurate than a balanced (positive sequence) one to inform a decision process; similarly using the exact nonlinear physics equations is more accurate than using a linear approximation.

Nevertheless, models are only useful if they state something that can be verified, i.e. make a statement about reality. Therefore, it is important to keep prioritising the link between reality and models, and let that link guide data collection standards and practices. It cannot be overstated that the physics of networks objectively characterise challenges and equity in network access. If a network is congested, it is the physics that determine which part of the networks experience congestion first, and which PV systems are curtailed first. Obviously, we have technologies (active filters, capacitor banks) that exploit the physics to open up additional network capacity as well – but again, these tools are made real by the physics of networks.

Due to the large amount of data, in propriety data formats, with a complex data model, the key challenges in this project were positioned at the intersection of data science and software engineering, with input from domain expertise required for overall integration. Resourcing such multidisciplinary projects is well-known to be challenging. It is noted that throughout the project meetings, DNSP employees have commented that they face the same challenges in their own organisations. We note that due to data sensitivities, the CSIRO team was also limited in its ability to contract people to fill software engineering and data science roles within the project.

7.6 Low-voltage network data sharing and privacy

Despite wide agreement on the need for greater data sharing between the electricity industry, researchers, and the public, sharing data in an effective manner remains a challenge, and this project was not immune to such hurdles. While it is unclear what was the amount of data CSIRO received compared to the relevant data that DNSPs engaged in the project have access to, in general, it was clear that many of the networks refrained from sharing significant amounts of useful data. While we can only speculate at the reasons for why some networks may have withheld data. It is likely related to privacy and/or security concerns, a lack of capability or resources at the DNSPs to facilitate data sharing even if they wanted to, and a concern about the risk of losing or sacrificing some strategic opportunity. Obviously, data sensitivities around customer data and location are reasonable. Nevertheless, separating customer data from network data largely resolves many of the immediate data sharing challenges – if tools to do so are available. In any case, planning engineers at DNSPs also often use synthetic or representative data for customer loads – instead of measured data - during the network design phase.

A partial solution would be to develop differential privacy-based network data cleaning tools (Fioretto, et al., 2020). Differential privacy can give mathematically provable guarantees about not leaking information relate to individual 'records' (Dwork & Roth, 2013). This could be used to obfuscate sensitive data while maintaining representativeness of the electrical engineering features. Continued research is necessary to extend this to distribution networks, while avoiding pitfalls and misapplication (Domingo-Ferrer, et al., 2020).

In general, significantly greater attention needs to be paid to data sharing – by researchers, networks, and policymakers – in order to overcome the hurdles identified above. Australia's electricity network businesses are granted a monopoly to provide an essential public service. As such, sharing their data with the broader public for legitimate uses – including developing new tools and methods to better plan and operate networks – should be prioritised, while ensuring that reasonable privacy and security concerns are recognised and accommodated. Improving data literacy within networks and the broader industry – including regulators, researchers, and technology providers – is an essential step in unlocking data sharing challenges. If nothing else, hopefully this project has helped highlight some of these data sharing problems and made a small, if meaningful impact on the digital literacy of those engaged.

8 Conclusions and Future Research

In order to improve the hosting capacity of distributed energy resources (DER) across Australia, a greater understanding is required of low-voltage networks, to allow design and assessment of appropriate technologies and systems.

The National Low-Voltage Feeder Taxonomy Study is the first national low-voltage network taxonomy that outlines the real-world characteristics of Australian low-voltage distribution systems.

In doing so, we have developed a realistic, publicly available, dataset and models describing the most common types of low-voltage networks found in Australia.

The purpose is to inform and facilitate early exploration in simulation of new energy technologies – it does not avoid the need for assessment for deployment in specific scenarios experienced in different distribution network areas. These models will enable users to test the value proposition of innovative technological solutions by highlighting how they contribute to the stability, reliability, and performance, of networks across Australia.

The project has also identified specific areas where we need to improve the understanding of the challenges and opportunities associated with using data to improve DER integration and encourage greater overall utilisation of the network.

The 23 identified representative network models plus associated workbooks can be downloaded now as: **'Low-Voltage Feeder Taxonomy Study - Data & models'** release hosted on the Australia's National Energy Analytics Research Program website (NEAR) website. Download via *https://near.csiro.au/assets/f325fb3c-2dcd-410c-97a8-e55dc68b8064* .

8.1 Challenges

The availability of appropriate-quality network data is crucial to enable the transformation of decision processes in power system operations and development. For distribution networks, more, better-quality, and more accessible, network data offers the opportunity to move on from a 'fit-and-forget' network management approach to an active approach. Different levels of network data quality enable different levels of functionality (Dubey, et al., 2020). High-quality network data is useful for simulation analysis of new energy technologies but becomes a necessity to perform state estimation and orchestration. We note that developing next-level network data sets is considered one of the first milestones to enable advanced distribution management applications[24].

1 [24]"How to Create an Accurate Network Model and Dynamic State Data for an Advanced Distribution Management System", Vahraz Zamani and Terry Nielsen, https://smartgrid.ieee.org/newsletters/april-2020/how-to-create-an-accurate-network-model-and-dynamic-state-data-for-an-advanced-distribution-management-system-adms?utm_source=sg-monthly-april2020&utm_medium=email&utm_campaign=2020-enewsletter

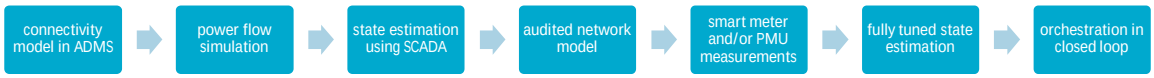


Figure 64 **Milestones in the deployment of ADMS, adapted from Vahraz Zamani and Terry Nielsen**

Arguably, most of the change that happens in distribution networks today is driven by the changes in the low-voltage aspects. Photovoltaic systems have become commonplace in residential public networks in Australia and other parts of the world. Newer technologies such as batteries and electric vehicles are set to become popular within the next ten years. Many of these new technologies are very flexible, and part of that flexibility could be used to support networks, without adversely affecting customers. Well-targeted network, power-electronic, and communication standards are needed to achieve this vision, while keeping costs under control and reliability high.

It is important that simulation data sets correspond to reality – the whole point of pursuing physics-based modelling is to make statements that are falsifiable and verifiable, i.e. inconsistencies can be resolved by measuring properties in the real world. The physics of power networks are well-understood, as are the algorithms to solve the resulting sets of equations, but the quality of existing network data is generally insufficient to develop best-practice simulation models (e.g. with explicit neutral conductor, with correct phase assignment of customers). While it would be impractical and costly to improve the data quality for all distribution grids in the near term, there are opportunities to improve the data quality over time, by designing better processes and workflows. We note that physics-based approaches to managing network access and operations have worked well for transmission network service providers and independent system operators. The physics ultimately describes what the level playing field is, and defines a way to create a consensus among all the stakeholders of the system limits, and how it affects those that are in congested parts.

Test beds in simulation can be used to inform the integration of new energy technologies. For Australia, the publicly available information on low-voltage networks was very limited before this project. As an outcome of this project, 23 data sets on LV networks were developed and released to the public. This data sets can be used to set up unbalanced power flow simulation studies. Illustrations of how such studies can be set up, have been released as ‘notebooks’, relying on OpenDSS as the unbalanced power flow simulation engine.

This project ran into a wide range of data quality challenges and is consequently limited in scope. Therefore, a broader exploration of network data quality issues and solutions is proposed in the future work section. Improving network data quality in the long-term is crucial to enable the development and deployment of novel automation technologies on advanced distribution management platforms (Claeys, et al., 2021; Dubey, et al., 2020). We specifically highlight the need for high-quality frameworks and tools that enable

- conversion between different network data formats;
- cleaning, debugging and maintenance of network data, while making sure it preserves the ability to run power flow solvers;

- obscuring privacy-sensitive data in power flow cases in an automated fashion, minimising leakage risks due to imperfect practice.

Platforms such as LFEnergy[25], RACE for Networks[26] and GlobalPST [27]could bring NSPs and research organisations together globally to make progress on such tools.

8.2 Future research avenues

The following is a list of research ideas coming from lessons learnt in the project:

- Clustering using power flow metrics combined with the network metrics
    - This will naturally separate clusters of congested and lightly load networks
- Foster the development and tools and workflows for
    - Network data cleaning, improving data quality, …
    - Setting up probabilistic power flow quickly, using public data sets.
    - Releasing privacy-sensitive network data sets
- Obtain larger data sets for Australia:
    - High time resolution active and reactive power measurements are needed to judge network losses more accurately
    - Voltage-dependence of different categories of LV loads
- Develop best practice documents around network data and model building
    - Validate the best practice in the real world by instrumenting a well-known network, to understand potential mismatches between simulation and reality better.
    - Analyse decision risks of using approximate models in difficult contexts
- Development of data models for distribution networks that are easier to check and validate automatically. Simulating real networks is preferable to using synthetic ones or combining synthetic ones. For the reasons discussed earlier, we believe the next step should be to develop integrated MV-LV data sets. In the long term, this needs to be integrated with HV as well.
    - The interactions between MV and LV can be significant and need to be understood better. For electric vehicle integration, integrated LV+MV power flow data sets may prove useful to identify congestion risks related to fast charging in MV grids and slow charging in LV grids.
- Developing user-friendly wrappers to make tools like OpenDSS(Direct) more easily usable and accessible to engineers at DNSPs, with design visions similar to PandaPower[28]. For

[25] "LF Energy is an open source foundation focused on the power systems sector, hosted within The Linux Foundation" https://www.lfenergy.org/

[26] https://www.racefor2030.com.au/race-for-networks/

[27] https://globalpst.org/

[28] http://www.pandapower.org/

example, in PandaPower, all elements can be defined with nameplate parameters and are internally processed with equivalent circuit models. If the engineer is not forced to do this manually, there are fewer opportunities for error, and the network data is easier to maintain and transform over time. Another useful functionality would be to develop convenience functions for element definition as well as standard component libraries to make it easy for the engineers to automate the creation of new network data sets.

9 References

Berry, A., Moore, T., Ward, J., Lindsay, S., Proctor, K., & Mitchell, T. (2013). *National Feeder Taxonomy: Describing a Representative Feeder Set for Australian Electricity Distribution Networks.* Newcastle: CSIRO.

Broderick, R. J., & Williams, J. R. (2013). Clustering Methodology for Classifying Distribution Feeders. *IEEE 39th Photovoltaic Specialists Conference (PVSC)*. doi:10.1109/PVSC.2013.6744473

Carson, J. T. (1926). Wave Propagation in Overhead Wires with Ground Return. *Bell System Technical Journal, 5*(4), 539-54. doi:10.1002/j.1538-7305.1926.tb00122.x

Claeys, S., Vanin, M., Geth, F., & Deconinck, G. (2021). Applications of optimization models for electricity distribution networks. *Wiley Interdisciplinary Reviews: Energy and Environment, e401*.

Crownshaw, T., Miller, A., Lemon, S., McNab, S., & Strahan, R. (2016). *Determination of Distributed Generation Hosting Capacity in Low-voltage Networks and Industry Applications.* Wellington: EEA Conference & Exhibition, 22 - 24 June.

Dale, M. (2013). *LV Network Templates for A Low-Carbon Future.* Bath: Western Power Distribution.

Dickert, J., Domagk, M., & Schegner, P. (2013). Benchmark Low Voltage Distribution Networks Based on Cluster Analysis of Actual Grid Properties. *IEEE Grenoble Conference*. doi:10.1109/PTC.2013.6652250

Dinning, A., Tan, D. S., Pradhan, P., Casey, S., Thiris, C., Williams, C., . . . Logan, S. (2020). *Future grid for distributed energy.* ENEA Consulting.

Domingo-Ferrer, J., Sánchez, D., & Blanco-Justicia, A. (2020). The Limits of Differential Privacy (and its Misuse in Data Release and Machine Learning). *Association for Computing Machinery, 1*(1), 1-4. Retrieved from https://arxiv.org/pdf/2011.02352.pdf

Dubey, A., Bose, A., Liu, M., & Ochoa, L. (2020). Paving the way for advanced distribution management systems applications: Making the most of models and data. *IEEE Power Energy Magazine, 18*, 63-75. doi:https://doi.org/10.1109/MPE.2019.2949442

Dwork, C., & Roth, A. (2013). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science, 9*(3-4), 211-487. doi:https://doi.org/10.1561/0400000042

EA Technology Ltd. (2018). *LV Management Strategy Annexe 2: Development of the Transform Model: 122250 Annexe 2.* SA Power Networks, 23 November .

Electricity North West. (2014). *Low Voltage Network Solutions Closedown Report.* Electricity Northwest. Retrieved from https://www.ofgem.gov.uk/system/files/docs/2017/04/lvns_closedown_report.pdf

Fioretto, F., Mak, T. W., & Van Hentenryck, P. (2020). Differential privacy for power grid obfuscation. *IEEE transactions on smart grid, 11*(2), 1356-66.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems, 17*, 107–145 . Retrieved from https://doi.org/10.1023/A:1012801612483

Jain, A., & Mather, B. (2018). Clustering Methods and Validation of Representative Distribution Feeders. *IEEE*.

Kapoor, S., Sturmberg, B., & Shaw, M. (est. 2020). *A Review of Publicly Available Energy Data Sets.* Australian National University.

Kersting, W. H., & Dugan, R. C. (2006). *Recommended Practices for Distribution System Analysis.*

Kersting, W. H., & Green, R. K. (2011). The application of Carson's equation to the steady-state analysis of distribution feeders. *IEEE/PES Power Systems Conference and Exposition* (pp. 1-6). Phoenix: IEEE.

Krause, O. (2019). *Solar Enablement Initiative: Final Report.* University of Queensland.

Levi, V., Strbac, G., & Allan, R. (2005). , 'Assessment of performance-driven investment strategies of distribution systems using reference networks,. *IEE Proceedings - Generation, Transmission and Distributio, 152*(1). doi:10.1049/ip-gtd:20041109

Levi, V., Strbac, G., & Allan, R. (2005). Assessment of performance-driven investment strategies of distribution systems using reference networks. *IEE Proceedings - Generation, Transmission and Distribution, 152*(1). doi:10.1049/ip-gtd:20041109

Li, F., & Shaddick, G. (2013). *Stresses on the LV Network caused by Low Carbon Technologies.* Bristol, BS2 0TB: Western Power Distribution (South Wales).

Li, F., & Shaddick, G. (2014). *Low voltage network solutions closedown report.* Warrington: Electricity Northwest.

Li, Y., & Wolfs, P. J. (2014). Taxonomic description for western Australian distribution medium-voltage and low-voltage feeders. *IET Generation, Transmission & Distribution, 8*(1), 104-113. doi:10.1049/iet-gtd.2013.0005

Ma, C. (2020). *A novel evaluation framework for energy losses in low voltage distribution grids.* Kassel, Germany: Kassel university press.

Ma, C., Menke, J.-H., Dasenbrock, J., Braun, M., Haslbeck, M., & Schmid, K.-H. (2019). Evaluation of energy losses in low voltage distribution grids with high penetration of distributed generation. *Applied Energy, 256*. doi:10.1016/j.apenergy.2019.113

Marcos, F., Domingo, C., Gómez San Román, T., Palmintier, B., Hodge, B., Krishnan, V., . . . Mather, B. (2017). A Review of Power Distribution Test Feeders in the United States and the Need for Synthetic Representative Networks. *Energies, 10*(11).

Navarro-Espinosa, A. (2014). *Deliverable 3.6: What-if Scenario Impact Studies based on real LV networks.* The University of Manchester.

Navarro-Espinosa, A. (2014). *Low Voltage Networks Models and Low Carbon Technology Profiles.* The University of Manchester. Retrieved from https://www.enwl.co.uk/globalassets/innovation/lvns/lvns-academic/summary-report.pdf

Navarro-Espinosa, A., & Ochoa, L. F. (2016). Probabilistic Impact Assessment of Low Carbon Technologies in LV Distribution Systems. *IEEE Transactions on Power Systems, 31*(3), 2192-2203. doi:10.1109/tpwrs.2015.2448663

Nijhuis, M., Gibescu, M., & Cobben, S. (2015, 15-18 June). Clustering Of Low Voltage Feeders Form A Network Planning Perspective. *presented at the 23rd International Conference on Electricity Distribution*.

Pecenak, Z. K., Disfani, V. R., Reno, M. J., & Kleissl, J. (2018). *IEEE TRANSACTIONS ON POWER SYSTEMS*, 1320-1328, VOL. 33, NO. 2, MARCH.

Procopiou, A. T., & Ochoa, L. N. (2019). *Advanced Planning of PV-Rich Distribution Networks - Deliverable 1: HV-LV modelling of selected HV feeders.* University of Melbourne.

Procopiou, A. T., & Ochoa, L. N. (2019). *Advanced Planning of PV-Rich Distribution Networks - Deliverable 2: Innovative Analytical Techniques.* The University of Melbourne.

Procopiou, A. T., Liu, M. Z., & Nacmanson, W. (2020). *Advanced Planning of PV-Rich Distribution Networks – Deliverable 4: Non-Traditional Solutions.* University of Melbourne.

Procopiou, A. T., Petrou, K., & Ochoa, L. N. (2020). *Advanced Planning of PV-Rich Distribution Networks – Deliverable 3: Traditional Solutions.* The University of Melbourne.

Ratnam, E., Weller, S., Kellett, C., & Murray, A. (2017). Residential load and rooftop PV generation: an Australian distribution network dataset. *International Journal of Sustainable Energy, 36*(8). doi:10.1080/14786451.2015.1100196

Rigoni, V., & Ochoa, L. (2014). *Deliverable 3.7: Characterisation of LV Networks.* Manchester: University of Manchester.

Rigoni, V., Ochoa, L. F., Chicco, G., Navarro-Espinosa, A., & Gozel, T. (2016). Representative residential LV feeders: A case study for the North West of England. *IEEE Transactions on Power Systems, 31*(1), 348–360.

Santos-Martin, D., & Lemon, S. (2016). Simplified Modeling of Low Voltage Distribution Networks for PV Voltage Impact Studies. *IEEE TRANSACTIONS ON SMART GRID, VOL. 7, NO. 4, JULY*, 1924-1931.

Schneider, K. P., Chen, Y., Engle, D., & Chassin, D. (2009). A Taxonomy of North American Radial Distribution Feeders. *IEEE Power & Energy Society General Meeting*.

Schneider, K. P., Engel, D. W., Chen, Y., Thompson, S. E., Chassin, D. P., & R. G. Pratt. (2008). *Modern Grid Initiative Distribution Taxonomy Final Report.* Pacific Northwest National Laboratory.

Shafiei, M., Liu, A., Ledwich, G., Walker, G., Morosini, G., & Terry, J. (2019). 'Solar Enablement Initiative in Australia: Report on Efficiently Identifying Critical Cases for Evaluating the Voltage Impact of Large PV Investment. *IEEE Power & Energy Society General Meeting (PESGM)*. doi:10.1109/PESGM40551.2019.8973794

Urquhart, A. J., & Thomson, M. (2015). Impacts of demand data time resolution on estimates of distribution system energy losses. *IEEE Transactions on Power Systems, 30*(3), 1483-1491. doi:https://doi.org/10.1109/TPWRS.2014.2349157

van der Maaten, L., Postma, E., & Herik, H. (2007). Dimensionality Reduction: A Comparative Review. Journal of Machine Learning Research. *Journal of Machine Learning Research, 10*.

Yildiz, B., Stringer, N., Heslop, S., Bruce, A., Heywood, P., Macgill, I., & Passey, R. (2020). *Voltage Analysis of the LV Distribution Network in the Australian National Electricity Market. May. https://prod-energycouncil.energy.slicedtech.com.au/lv-vo.* Retrieved from https://prod-energycouncil.energy.slicedtech.com.au/lv-voltage-report

Appendix : Additional information

Appendix A Network Feature Descriptions

Table 17 shows the full list of metrics produced for each LV network model. The source column denotes whether the metric was developed as part of this project (LVFT) or came from the DiTTo code ('DiTTo'). Descriptions of the DiTTo metrics are available at https://nrel.github.io/ditto/metrics/. LVFT metrics are described below. Repeated or obvious descriptions are omitted for brevity. Not all metrics were available for all networks, and units for some measurements differed between DNSPs (e.g. wire diameters) and were not normalised.

Table 17 List of all network features extracted for each LV network model

Feature	Source	Description
dnsp	LVFT	DNSP Name
name	LVFT	Network name
avg_lat	LVFT	Average latitude for network
avg_long	LVFT	Average longitude for network
n_lat_long	LVFT	Number of lat/long points used for averages
mb_code_2016	LVFT	Meshblock code for (avg_lat, avg_long)
mb_category_name	LVFT	Meshblock name
sa2_name	LVFT	Meshblock Statistical Area Level 2
sa3_name	LVFT	Meshblock Statistical Area Level 3
sa4_name	LVFT	Meshblock Statistical Area Level 4
ggcsa_name	LVFT	Greater Capital City Statistical Areas (GCCSA)
state_name	LVFT	State name
ra_code_2016	LVFT	Remoteness Area Code
ra_name_2016	LVFT	Remoteness Area Name
n_line_types	LVFT	Number of distinct line types

Feature	Source	Description
cycles_removed	LVFT	Number of graph cycles removed from the network
n_PowerTransformer	LVFT	Number of power transformers
n_Winding	LVFT	Number of transformer windings
n_PhaseWinding	LVFT	Number of transformer phase windings
n_Regulator	LVFT	Number of regulators
n_Line	LVFT	Number of lines
n_Wire	LVFT	Number of wires
n_Node	LVFT	Number of nodes
n_Position	LVFT	Number of positions
n_Load	LVFT	Number of loads
n_PhaseLoad	LVFT	Number of phaseloads
n_PowerSource	LVFT	Number of power sources
n_duplicate_names	LVFT	Number of duplicate element names in the parsed model
n_disconnected_graphs	LVFT	Number of disconnected graphs found in the parsed model
n_cycles	LVFT	Number of remaining graph cycles
total_line_length	LVFT	Total length of lines
n_edges	LVFT	Number of graph edges
n_nodes	LVFT	Number of graph nodes (any non-edge element)
n_deg_1_node	LVFT	Number of degree-1 nodes
n_deg_2_node	LVFT	
n_deg_3_node	LVFT	

Feature	Source	Description
n_deg_4_node	LVFT	
n_deg_>4_node	LVFT	Number of > degree-4 nodes
n_1_wire_lines	LVFT	Number of 1-wire lines
n_2_wire_lines	LVFT	Number of 2-wire lines
n_3_wire_lines	LVFT	Number of 3-wire lines
max_hops_to_sub	LVFT	Maximum hops (edges) from any node to substation
min_hops_to_sub	LVFT	Minimum hops (edges) from any node to substation
mean_hops_to_sub	LVFT	Mean number of hops (edges) for all paths from nodes to substation
median_hops_to_sub	LVFT	Median hops (edges) ‘
min_dist_to_sub	LVFT	Minimum distance (m) from all nodes to substation
max_dist_to_sub	LVFT	
mean_dist_to_sub	LVFT	
median_dist_to_sub	LVFT	
min_R0_to_sub	LVFT	Minimum R0 from all nodes to substation
max_R0_to_sub	LVFT	
mean_R0_to_sub	LVFT	
median_R0_to_sub	LVFT	
min_R1_to_sub	LVFT	
max_R1_to_sub	LVFT	
mean_R1_to_sub	LVFT	
median_R1_to_sub	LVFT	

Feature	Source	Description
min_X0_to_sub	LVFT	
max_X0_to_sub	LVFT	
mean_X0_to_sub	LVFT	
median_X0_to_sub	LVFT	
min_X1_to_sub	LVFT	
max_X1_to_sub	LVFT	
mean_X1_to_sub	LVFT	
median_X1_to_sub	LVFT	
n_loads_240V	LVFT	Number of ~240v loads
n_loads_400V	LVFT	Number of ~400v loads
n_loads_>400V	LVFT	Number of ~>400v loads
n_lines_240V	LVFT	
n_lines_400V	LVFT	
n_lines_>400V	LVFT	
n_nodes_240V	LVFT	
n_nodes_400V	LVFT	
n_nodes_>400V	LVFT	
powersource_rated_power	LVFT	Rated powersource power
powersource_n_phases	LVFT	Number of powersource phases
min_wire_radius	LVFT	Minimum wire radius (units un-normalised & dependant on source data)
max_wire_radius	LVFT	
mean_wire_radius	LVFT	

Feature	Source	Description
min_wire_diameter	LVFT	
max_wire_diameter	LVFT	
mean_wire_diameter	LVFT	
n_A_phase_wires	LVFT	Number of A-phase wires
n_B_phase_wires	LVFT	
n_C_phase_wires	LVFT	
n_overhead_lines	LVFT	
n_underground_lines	LVFT	
ratio_lines_overhead	LVFT	ratio of overhead-to-underground lines
n_feeders	LVFT	Number of feeders identified with heuristic
feeder_total_line_length	LVFT	Total length (m) of lines in feeders
feeder_total_n_edges	LVFT	
feeder_total_n_nodes	LVFT	
feeder_avg_total_line_length	LVFT	
feeder_avg_n_edges	LVFT	
feeder_avg_n_nodes	LVFT	
feeder_avg_n_deg_1_node	LVFT	
feeder_avg_n_deg_2_node	LVFT	
feeder_avg_n_deg_3_node	LVFT	
feeder_avg_n_deg_4_node	LVFT	
feeder_avg_n_deg_>4_node	LVFT	
feeder_avg_n_1_wire_lines	LVFT	
feeder_avg_n_2_wire_lines	LVFT	

Feature	Source	Description
feeder_avg_max_hops_to_sub	LVFT	
feeder_avg_min_hops_to_sub	LVFT	
feeder_avg_mean_hops_to_sub	LVFT	
feeder_avg_median_hops_to_sub	LVFT	
feeder_avg_min_dist_to_sub	LVFT	
feeder_avg_max_dist_to_sub	LVFT	
feeder_avg_mean_dist_to_sub	LVFT	
feeder_avg_median_dist_to_sub	LVFT	
feeder_avg_min_R0_to_sub	LVFT	
feeder_avg_max_R0_to_sub	LVFT	
feeder_avg_mean_R0_to_sub	LVFT	
feeder_avg_median_R0_to_sub	LVFT	
feeder_avg_min_R1_to_sub	LVFT	
feeder_avg_max_R1_to_sub	LVFT	
feeder_avg_mean_R1_to_sub	LVFT	
feeder_avg_median_R1_to_sub	LVFT	
feeder_avg_min_X0_to_sub	LVFT	
feeder_avg_max_X0_to_sub	LVFT	
feeder_avg_mean_X0_to_sub	LVFT	
feeder_avg_median_X0_to_sub	LVFT	
feeder_avg_min_X1_to_sub	LVFT	
feeder_avg_max_X1_to_sub	LVFT	
feeder_avg_mean_X1_to_sub	LVFT	

Feature	Source	Description
feeder_avg_median_X1_to_sub	LVFT	
feeder_avg_n_3_wire_lines	LVFT	
num_regulators	DiTTo	See https://nrel.github.io/ditto/metrics/ for description of all DiTTo metrics below
num_fuses	DiTTo	
num_switches	DiTTo	
num_reclosers	DiTTo	
num_breakers	DiTTo	
num_capacitors	DiTTo	
num_sectionalisers	DiTTo	
num_customers	DiTTo	
num_links_adjacent_feeders	DiTTo	
num_overloaded_transformers	DiTTo	
num_distribution_transformers	DiTTo	
max_len_secondaries_mi	DiTTo	
sum_distribution_transformer_mva	DiTTo	
num_1ph_transformers	DiTTo	
num_3ph_transformers	DiTTo	
ratio_1ph_to_3ph_transformers	DiTTo	
avg_degree	DiTTo	
diameter	DiTTo	
avg_path_len	DiTTo	
avg_regulator_sub_distance_mi	DiTTo	

Feature	Source	Description
avg_capacitor_sub_distance_mi	DiTTo	
avg_recloser_sub_distance_mi	DiTTo	
max_sub_node_distance_mi	DiTTo	
num_loops	DiTTo	
lv_len_mi	DiTTo	
mv_len_mi	DiTTo	
mv_1ph_len_mi	DiTTo	
mv_oh_1ph_len_mi	DiTTo	
mv_2ph_len_mi	DiTTo	
mv_oh_2ph_len_mi	DiTTo	
mv_3ph_len_mi	DiTTo	
mv_oh_3ph_len_mi	DiTTo	
lv_1ph_len_mi	DiTTo	
lv_oh_1ph_len_mi	DiTTo	
lv_2ph_len_mi	DiTTo	
lv_oh_2ph_len_mi	DiTTo	
lv_3ph_len_mi	DiTTo	
lv_oh_3ph_len_mi	DiTTo	
sum_load_kw	DiTTo	
sum_load_pha_kw	DiTTo	
sum_load_phb_kw	DiTTo	
sum_load_phc_kw	DiTTo	
sum_load_kvar	DiTTo	

Feature	Source	Description
num_lv_1ph_loads	DiTTo	
num_lv_3ph_loads	DiTTo	
num_mv_3ph_loads	DiTTo	
perct_lv_pha_load_kw	DiTTo	
perct_lv_phb_load_kw	DiTTo	
perct_lv_phc_load_kw	DiTTo	
sum_lv_pha_load_kw	DiTTo	
sum_lv_phb_load_kw	DiTTo	
sum_lv_phc_load_kw	DiTTo	
avg_num_load_per_transformer	DiTTo	
num_load_per_transformer	DiTTo	
num_customer_per_transformer	DiTTo	
transformer_kva_distribution_0	DiTTo	
ratio_load_kW_to_transformer_KVA_distribution	DiTTo	
nominal_voltages_0	DiTTo	
nominal_voltages_1	DiTTo	
nominal_voltages_2	DiTTo	
nominal_voltages_3	DiTTo	
nominal_voltages_4	DiTTo	
nominal_voltages_5	DiTTo	
convex_hull_area_sqmi	DiTTo	
substation_name	DiTTo	
Feeder_type	DiTTo	

Feature	Source	Description
ratio_mv_len_to_num_cust	DiTTo	
perct_mv_oh_len	DiTTo	
perct_lv_oh_len	DiTTo	
num_sectionalisers_per_recloser	DiTTo	
avg_load_pf	DiTTo	
avg_load_imbalance_by_phase	DiTTo	
nominal_medium_voltage_class	DiTTo	
cust_density	DiTTo	
load_density_kw	DiTTo	
load_density_kvar	DiTTo	
kva_density	DiTTo	
ditto_metrics_error	DiTTo	Error message from DiTTo if calculations failed

Appendix B Customer Load data sources

Data source	Data type	Location within Australia	Location detail	circuit level	Sampling Time	Data Length	Year Range	No of samples
Australian Low Energy Houses (LEH)	PV, load	Yes	Lochiel Park (LP) Green Village in Adelaide, South Australia	sub-circuit	1 min	1 years	2013	60
PVOutput	PV	Yes	customer across Australia	individual	5 min	5 years	2010 – 2020	5500
UMass Trace – Solar panels dataset	PV	No	Massachusetts, US	individual	1 min	1 years	2015	50
Ausgrid Solar home electricity data	PV, load	Yes	Ausgrid Zone	individual	30 min	3 years	2010 – 2013	300
Smart-Grid Smart-City	PV, load	Yes	Ausgrid Zone	individual	30 min	2 years	2010 – 2014	78200
MAISY	PV, load	No	Massachusetts, US	individual	15 min	1 years	1999 – 2020	7 000 000
Residential Building Energy Efficiency Study	PV, load	Yes	Melbourne Brisbane and Adelaide	sub-circuit	30 min	5 years	2013 – 2017	163
Reward Based Tariff (RBT)	Load	Yes	Energex and Ergon Zone	individual	30 min	2 years	2011 – 2013	504
UMass Trace – Apartment dataset	Load	No	Massachusetts, US	individual	15 min	2 years	2014 – 2016	114
Common Property Loads in Apartment Buildings	Load	Yes	Sydney	sub-circuit	15 min	1.5 years	Unknown	25
Literature: Spatio-temporal modelling of electric vehicle charging demand and impacts on peak household electrical load	Load	Yes	Victoria	individual	1 hour	1 years	2033	250
Public light profile From Appalachian Power,	Load	No	US	individual	1 hour	1 years	2016	Unclear

subsidiary of American Electric Power								
High-resolution Industrial Production Energy (HIPE) data	Load	No	Germany	individual	5.54 second	3 months	2017 – 2018	11
Household EnergyEnd-use Project (HEEP)	PV, load	No	New Zealand	sub-circuit	10 min	11 months	1999 – 2005	100
One-minute Solar irradiance	other	Yes	Across Australia	area aggregated	1 min	23 years	1997 – 2020	19
Original Zone Substation Load Data (NEAR Program)	Load	Yes	NSW: Ausgrid, Endeavour Energy ; VIC: CitiPower and Powercor, Jemena, United Energy; ACT: ActewAGL SA: SA Power Networks	substation	30 min	10 years	2008 - 2020	500
AEMO	Load	Yes	Victoria, NSW, ACT, SA, QLD and TAS	area aggregated	30 min	18 years	2002 - 2020	21
Standardised Zone Substation Gross Solar PV Generation (NEAR Program)	PV	Yes	Zone by DNSP in Australia	substation	1 hour	4 years	2011 - 2015	500
Small generation unit (SGU) installations	other	Yes	Across Australia	area aggregated	1 month	2 years	2019 - 2020	2000
UK Power Networks Zone	Load	No	UK Power Networks Zone (includes London, Cambridge, Brighton & Dover)	substation	10 min	5 months	2014	20
UMass Trace – Home dataset	PV, load	No	Massachusetts, US	sub-circuit	1 min	1 years	2016	6
UK Power Networks Zone	PV, load	No	UK Power Networks Zone (includes London, Cambridge, Brighton and Dover)	individual	1 min	3 months	2014	6
The Reference Energy Disaggregation Data Set	PV, load	No	Massachusetts, US	sub-circuit	3 second	3 weeks	2011	6

Table 18: PV and load data sources (1)

Data source	Measure-ment	Data Quality	Reason	Remark and Future Work	Data Accessibility
Australian Low Energy Houses (LEH)	Active Power	high	High data Quality (University of South Australia) strict selection of houses and installation of monitoring	Study done by University of South Australia Specific data for LEH in LP Green Village in Adelaide. Contact authors of the study for raw data	contact owner
PVOutput	Active Power	unknown		available data varies greatly by customers, ranging from 1 to >10 years	contact owner
UMass Trace – Solar panels dataset	Active Power	high	PV generation by solar panel		immediate download
Ausgrid Solar home electricity data	Active Power of general & controlled load, PV generation	high	DNSP	300 customers selected from a pool of 15000 with electricity consumption and PV generation of top and bottom 10% customers excluded. Details see section 2 article.	immediate download
Smart-Grid Smart-City	Active Power of general & controlled load, PV generation	high	DNSP & Aus gov.	Details of households, e.g. usage of gas heating, air-conditioning and clothes dryers, for some customers. One of the few load/PV profile in Australia. Can be correlated with electricity consumption	immediate download
MAISY				0.5s – 5mins. kW for project-based customers. Contact the company for free sample of data/trial of built-in visualisation tool, and validation of record of data	payment required
Residential Building Energy Efficiency Study	Active Power, gas and temperature	high	CSIRO project	Study of 209 houses (June 2012 to February 2013) Study of 163 houses (2013-2017) Specific data obtained by CSIRO, contact author of the study for data of metering	contact CSIRO
Reward Based Tariff (RBT)	active power of general and controlled load, PV generation	high	DNSP	Data obtained by CSIRO previously in this literature (p.18 table 2). The RBT scheme has used tariff to facilitate lower peak consumption, load profile data may be biased.	contact CSIRO

UMass Trace – Apartment dataset	Active Power	high	High data quality, With good documentation	With weather data	immediate download
Common Property Loads in Apartment Buildings	Active power of common property load	high	UNSW study	Study by UNSW which acquires CP loads data of apartments. Table 1 of the study summarise the characteristic of data of each building	contact owner
Literature: Spatio-temporal modelling of electric vehicle charging demand and impacts on peak household electrical load	SIMULATED House's Load profiles (active power) with EVs charging/discharging pattern	high	CSIRO project	Simulation with 4 variables: -EV uptake: data from CSIRO lit., 3-month interval -EV travel: data from VISTA and VicRoad obtained from Victoria government, 1-hour interval -Household energy and power: simulation of weather and calculation of cooling and heating load from a commercial engine, 1-hour interval -EV charging and discharging: data=EV travel model, assumed charging at home, 1-hour interval	contact CSIRO
Public light profile From Appalachian Power, subsidiary of American Electric Power	Active power of public light energy consumption	unknown	DNSP of US but unknown processing	datasheet of some low-energy usage utility equipment, e.g. public lightning, CCTV can be found in AEMO website.	immediate download
High-resolution Industrial Production Energy (HIPE) data	Active/Reactive Power, THD, voltage/current	high	Karlsruhe Institute of Technology, public research university	Load profiles of 11 machines for an electronics factory, see their publication	immediate download
Household EnergyEnd-use Project (HEEP)	Active Power	high	company BRANZ	Location and type of measurement see sections 1.1 and 4.1 respectively of the project report	contact owner
One-minute Solar irradiance	Solar irradiance (avg, min, max, std, ...)	high	Australia Bureau of Meteorology	1-minute data is processed instead of real time	immediate download
Original Zone Substation Load Data (NEAR Program)	Active Power, some with reactive power	high	DNSP	Request data from DNSP of the missing states at Australia Energy Regulator website	immediate download
AEMO	Active power of general and controlled load	high	AEMO	13 Net System Load Profile (NSLP), 8 Controlled Load Profile (CLP)	immediate download
Standardised Zone Substation Gross Solar	Active Power	unknown	methodology, field names unknown		immediate download

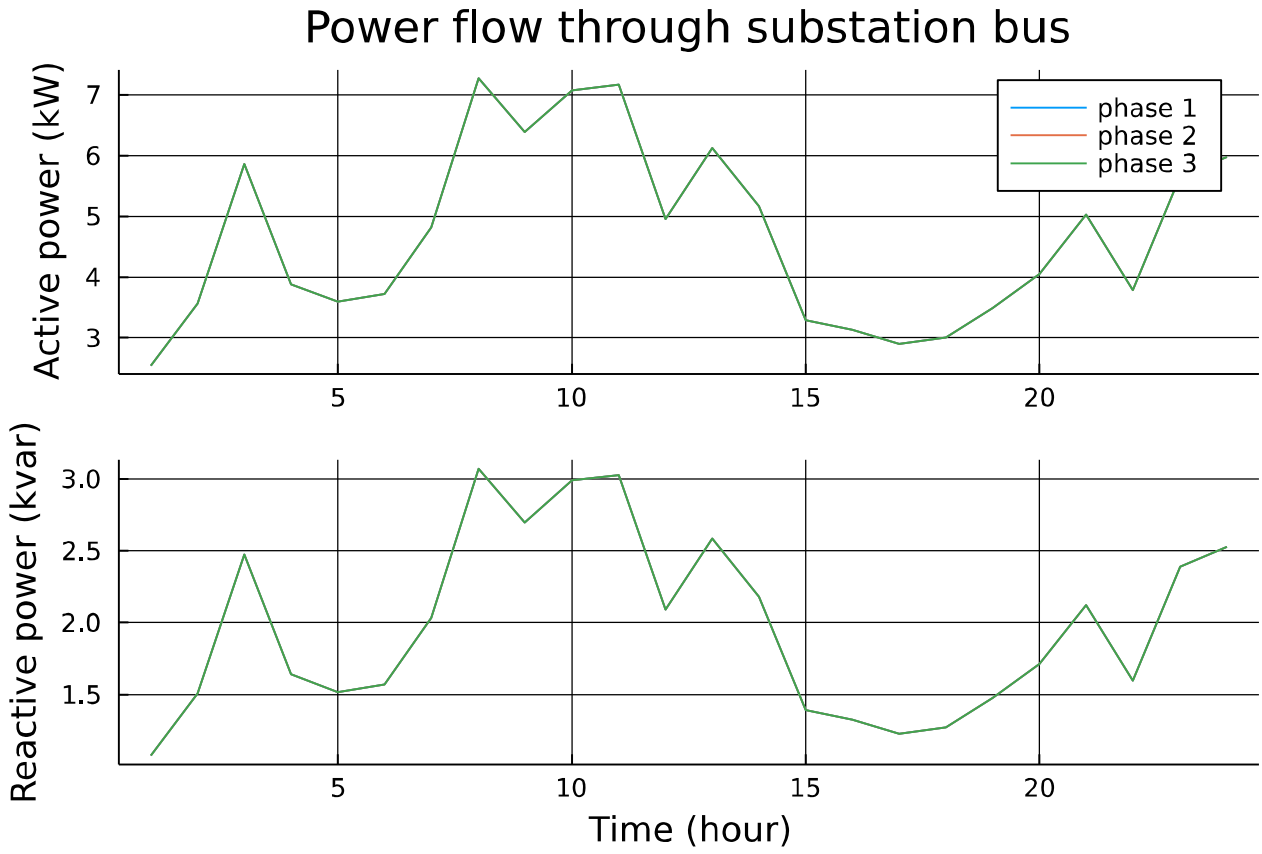
PV Generation (NEAR Program)					
Small generation unit (SGU) installations	small-scale PV installations by month and postcode	high	Australia Clean energy regulator		immediate download
UK Power Networks Zone	voltages at three phases	high	DNSP in UK		immediate download
UMass Trace – Home dataset	Active Power	high	Project based with documentation		immediate download
UK Power Networks Zone	Active/ Reactive Power, THD, voltage and current.	high	DNSP in UK	1-min and 10-min data for customers, feeders and network endpoints from mid-June 2014 to early Sep 2014. Hourly data for Oct 2013 to Oct 2014. Weather for the period of recording also provided. Detail data and well documented.	immediate download
The Reference Energy Disaggregation Data Set	Active power, reactive from AC waveform	high	publication of analysed data	Analysis of data see the publication	contact owner

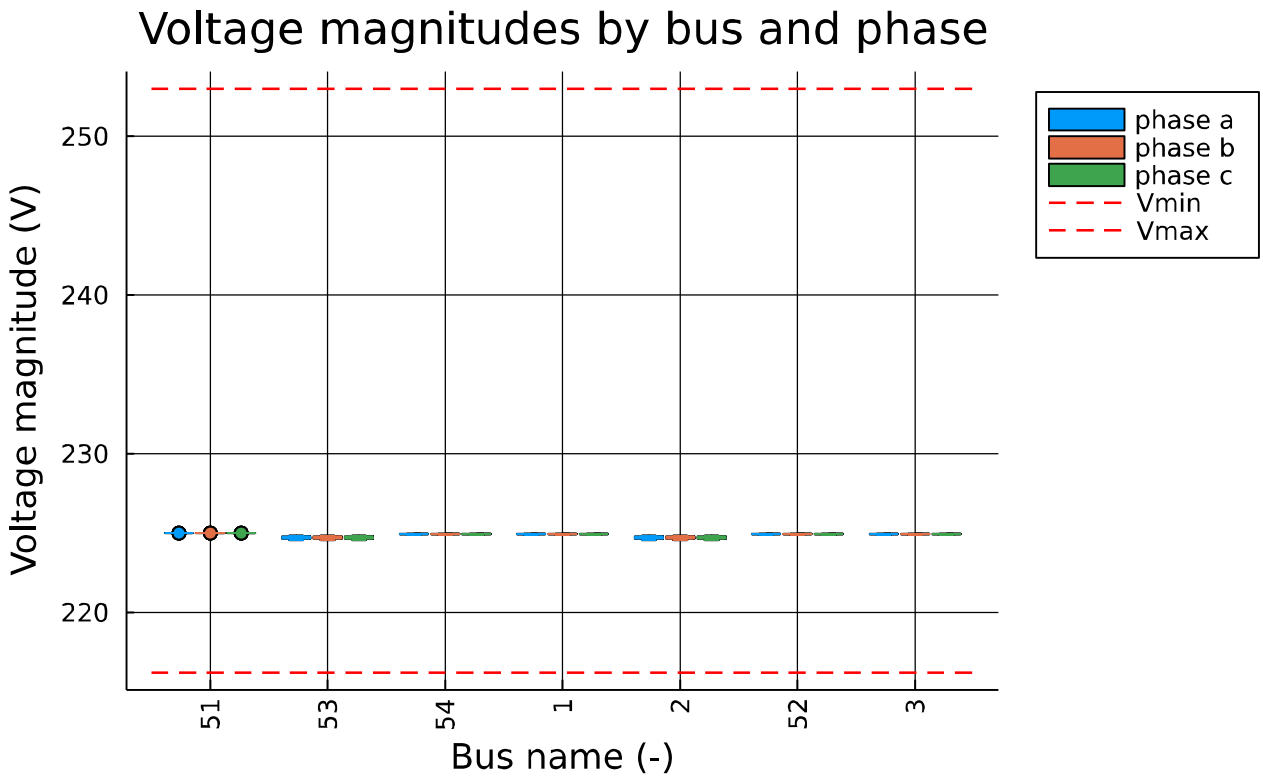
Table 19: PV and load data sources (2)

Appendix C Supplemental Power flow results – base cases

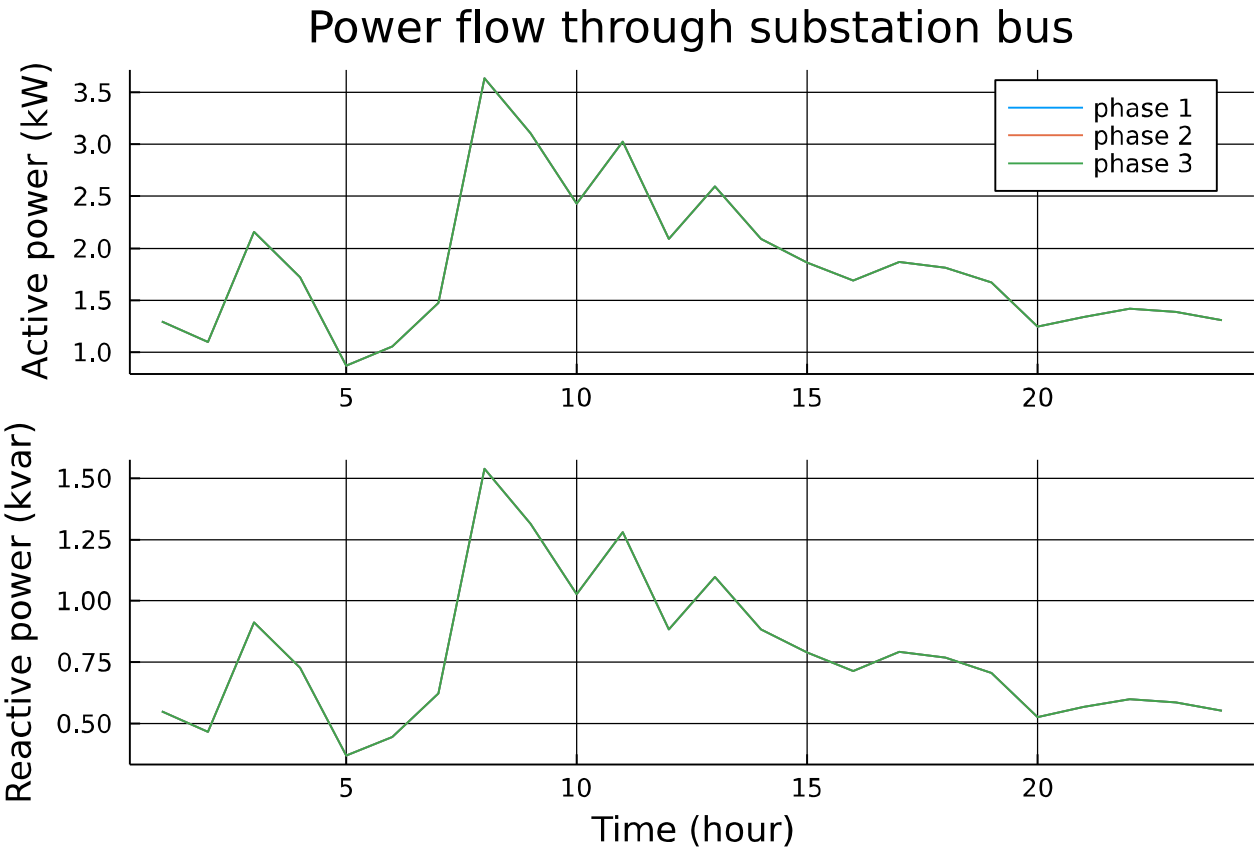
The body of the report only discusses a subset of the networks, i.e. those with interesting features. For completeness, we add the remaining base case results in this section.

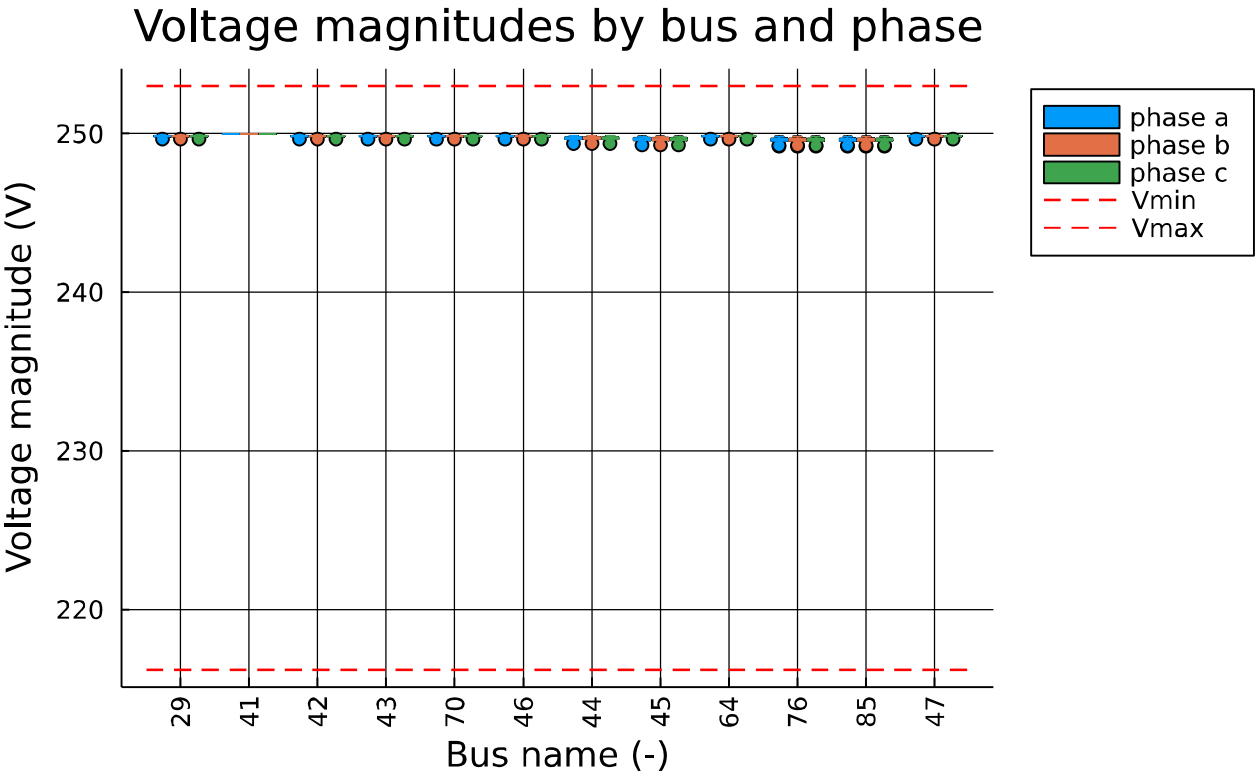
C.0 Network B


Power flow through substation bus
Active power (kW)
7
6
5
4
3
5
10
15
20
phase 1
phase 2
phase 3
Reactive power (kvar)
3.0
2.5
2.0
1.5
5
10
15
20
Time (hour)

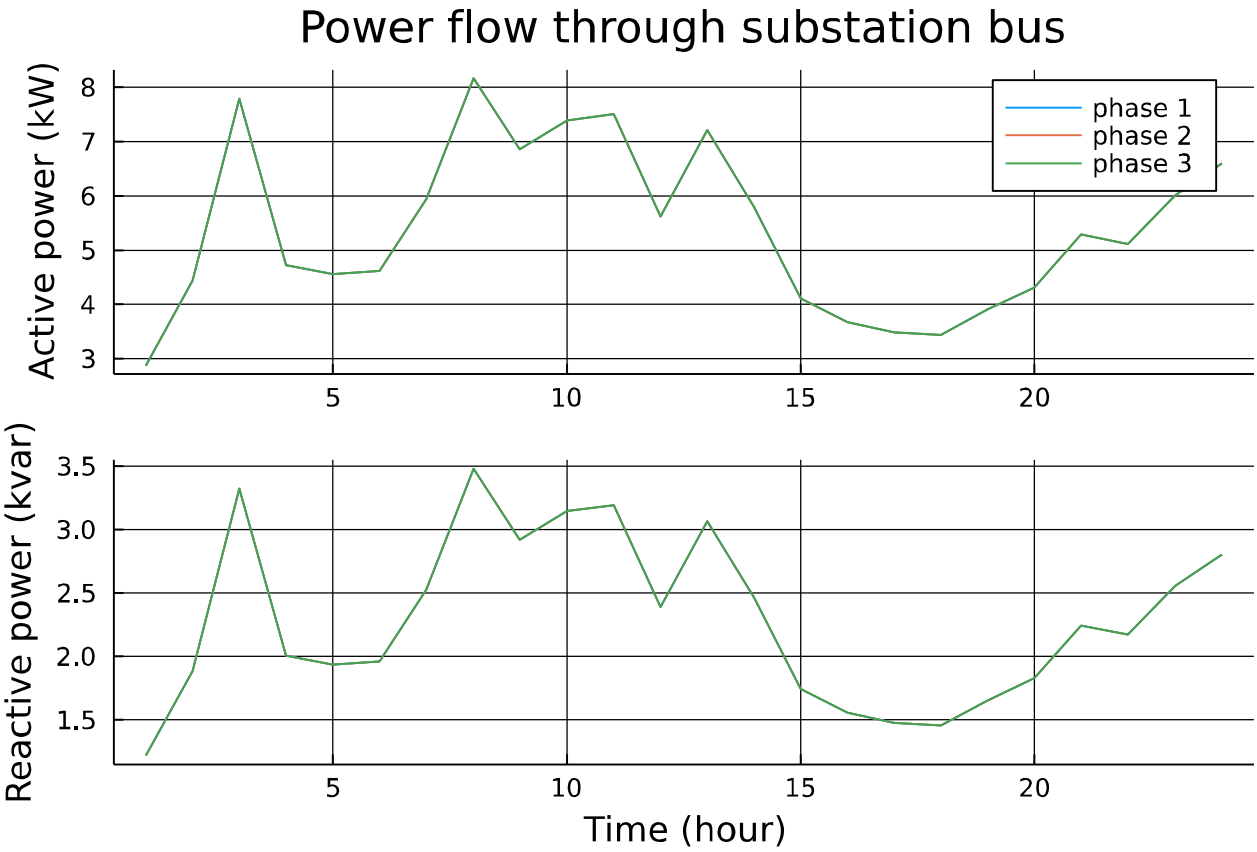

Voltage magnitudes by bus and phase
Voltage magnitude (V)
250
240
230
220
51
53
54
1
2
52
3
Bus name (-)
phase a
phase b
phase c
Vmin
Vmax

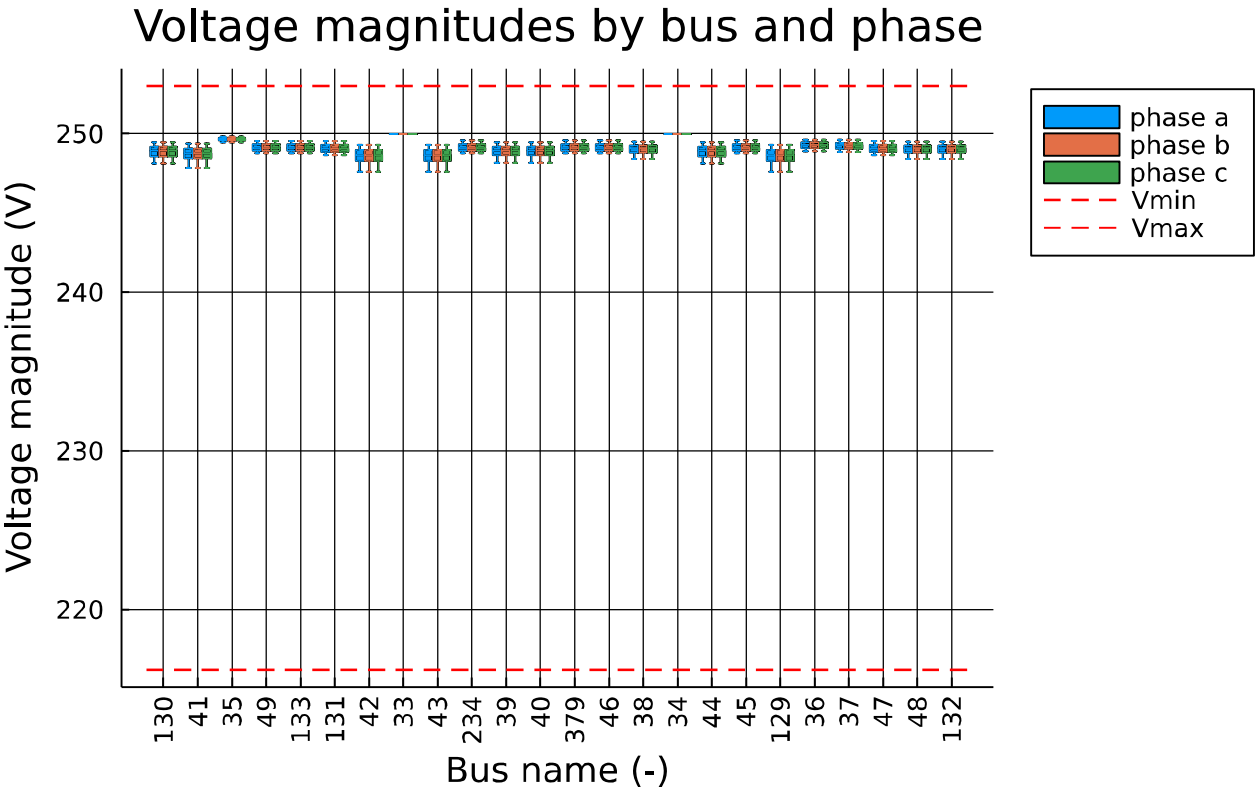
C.1 Network C


Power flow through substation bus
Active power (kW)
3.5
3.0
2.5
2.0
1.5
1.0
5
10
15
20
phase 1
phase 2
phase 3
Reactive power (kvar)
1.50
1.25
1.00
0.75
0.50
5
10
15
20
Time (hour)

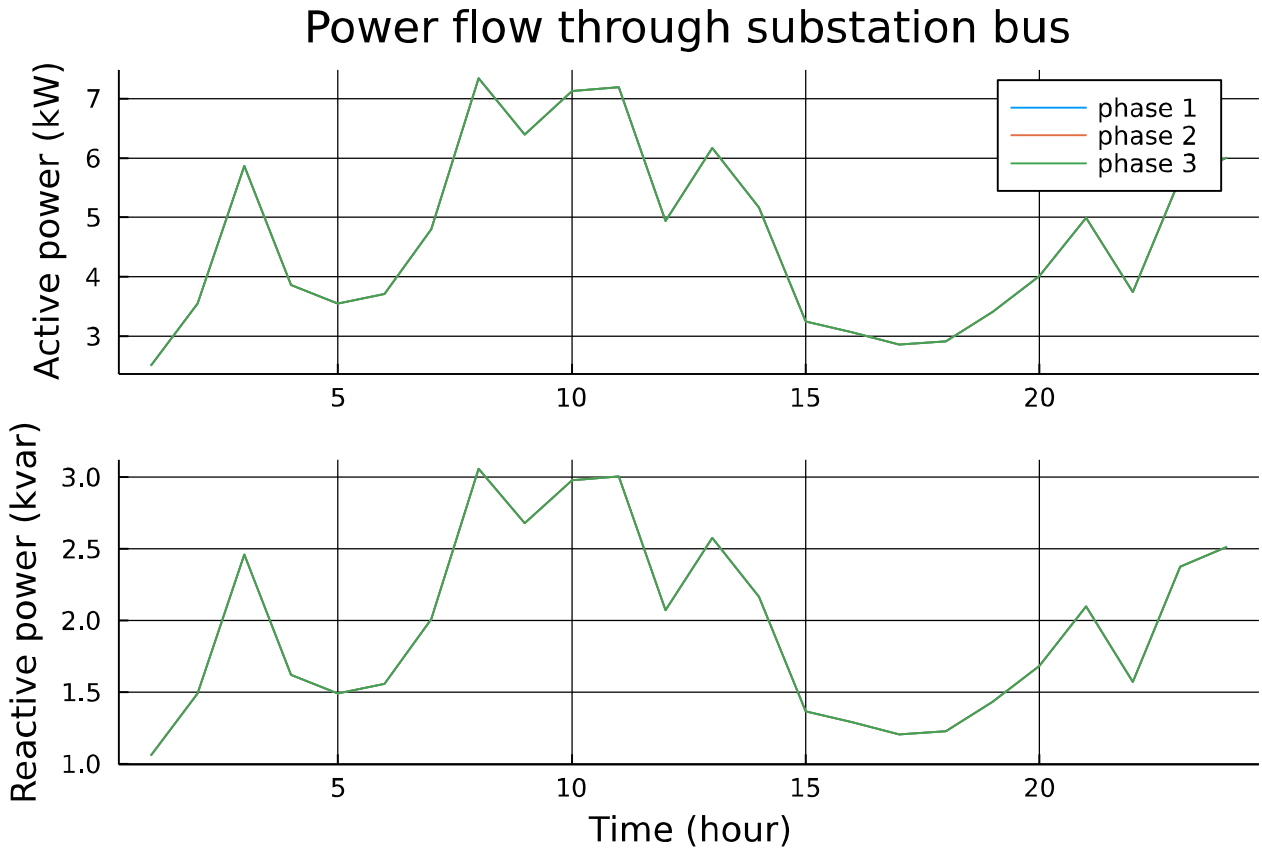

Voltage magnitudes by bus and phase
Voltage magnitude (V)
250
240
230
220
29
41
42
43
70
46
44
45
64
76
85
47
Bus name (-)
phase a
phase b
phase c
Vmin
Vmax

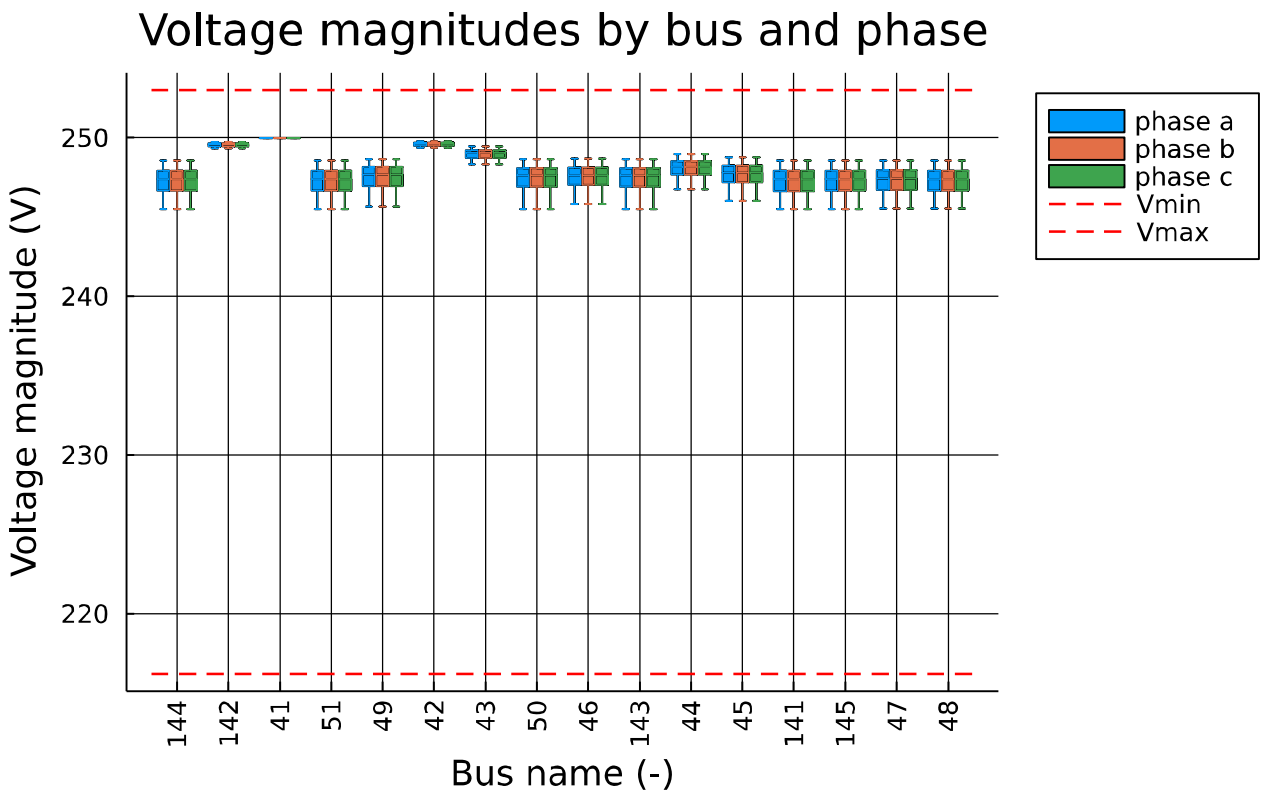
C.2 Network E


Power flow through substation bus
Active power (kW)
8
7
6
5
4
3
5
10
15
20
phase 1
phase 2
phase 3
Reactive power (kvar)
3.5
3.0
2.5
2.0
1.5
5
10
15
20
Time (hour)

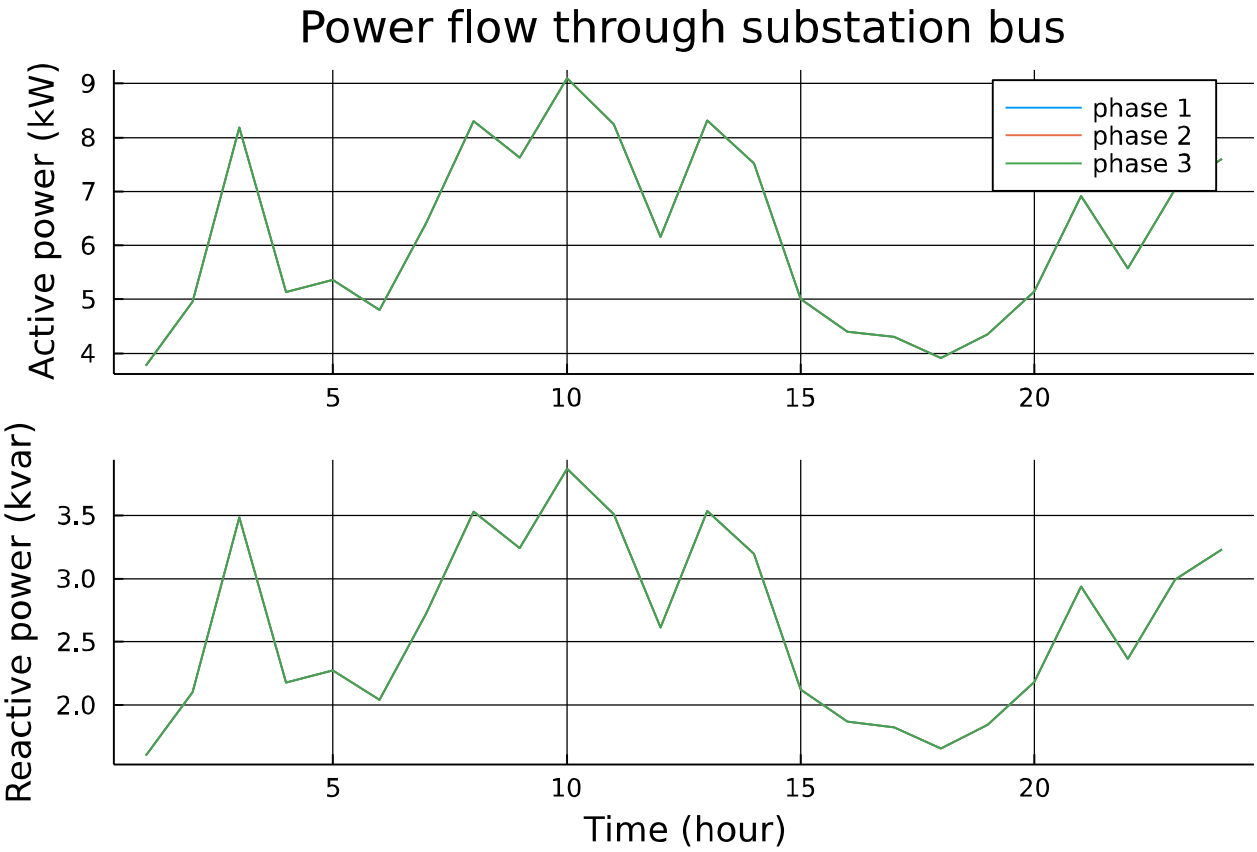

Voltage magnitudes by bus and phase
Voltage magnitude (V)
250
240
230
220
130
41
35
49
133
131
42
33
43
234
39
40
379
46
38
34
44
45
129
36
37
47
48
132
Bus name (-)
phase a
phase b
phase c
Vmin
Vmax

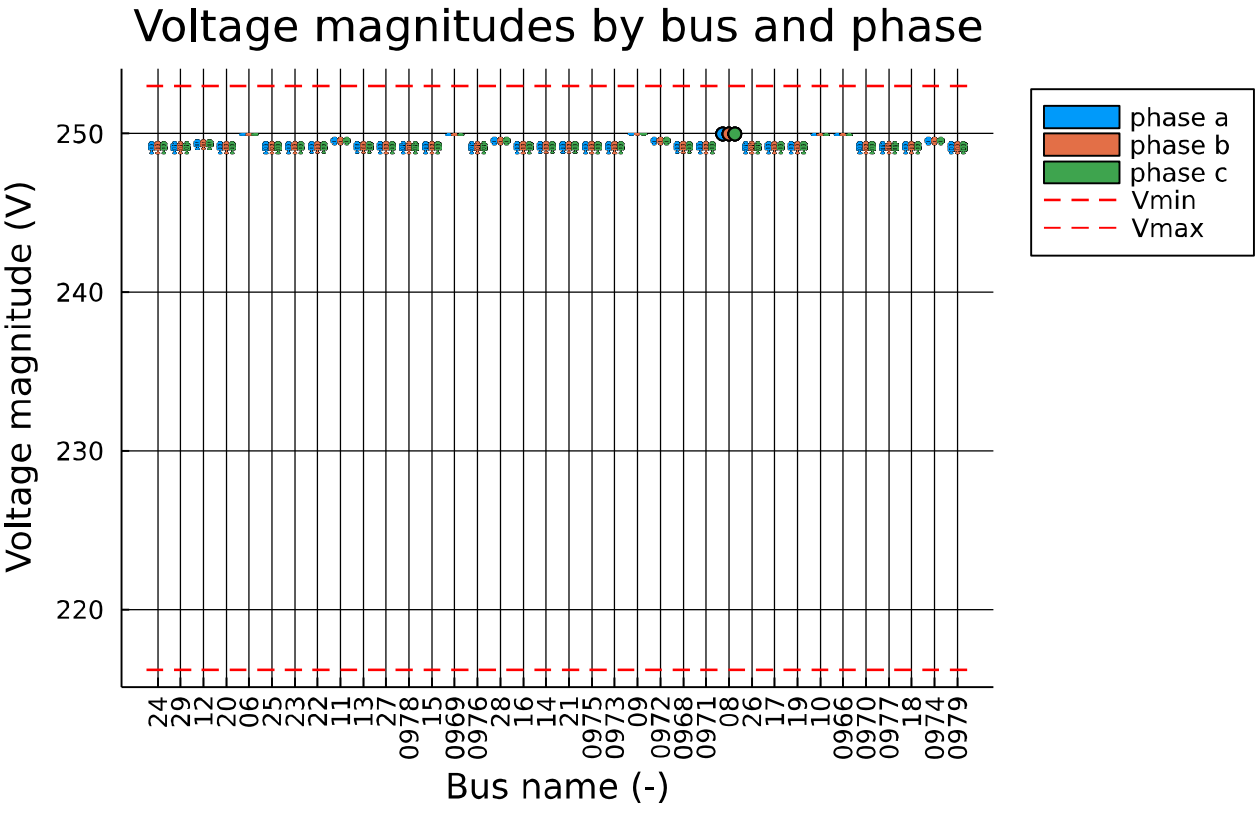
C.3 Network F


Power flow through substation bus
Active power (kW)
7
6
5
4
3
5
10
15
20
phase 1
phase 2
phase 3
Reactive power (kvar)
3.0
2.5
2.0
1.5
1.0
5
10
15
20
Time (hour)

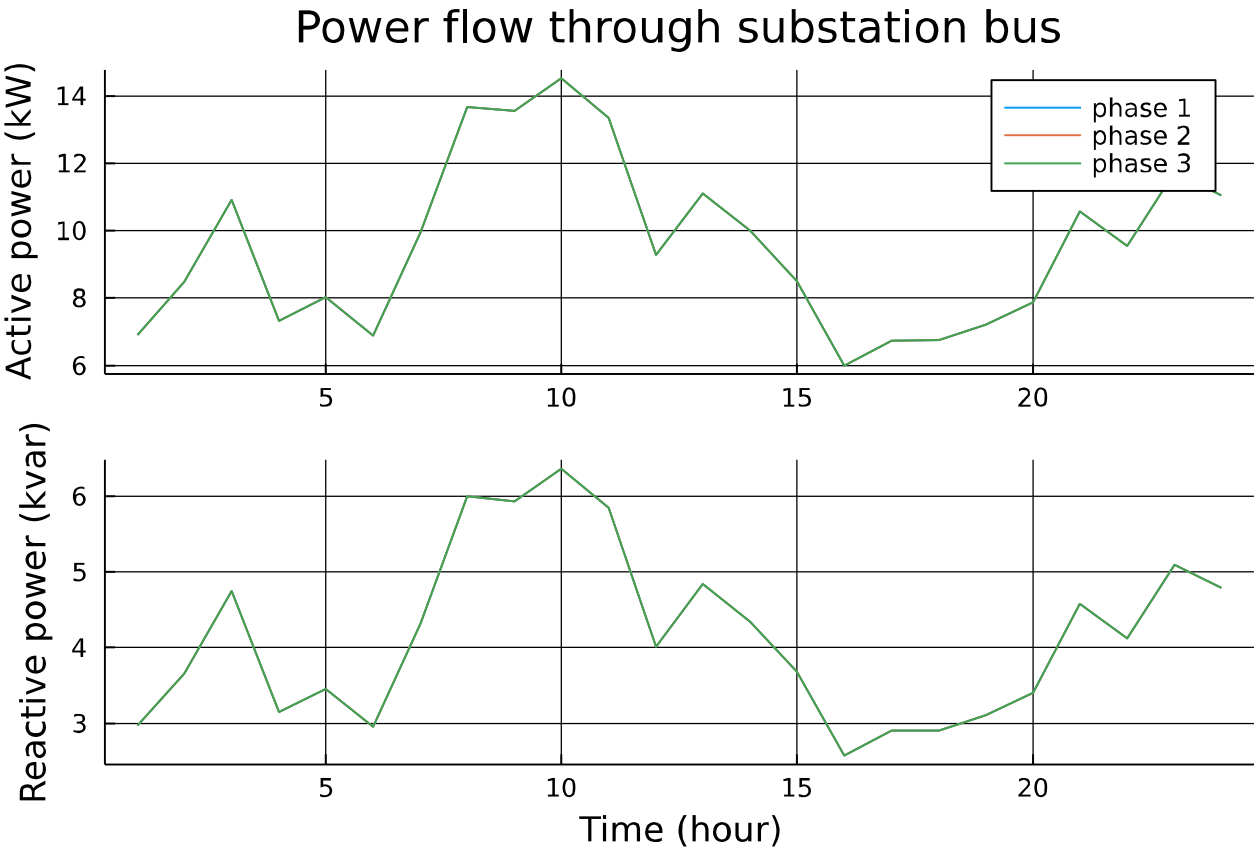

Voltage magnitudes by bus and phase
Voltage magnitude (V)
250
240
230
220
144
142
41
51
49
42
43
50
46
143
44
45
141
145
47
48
Bus name (-)
phase a
phase b
phase c
Vmin
Vmax

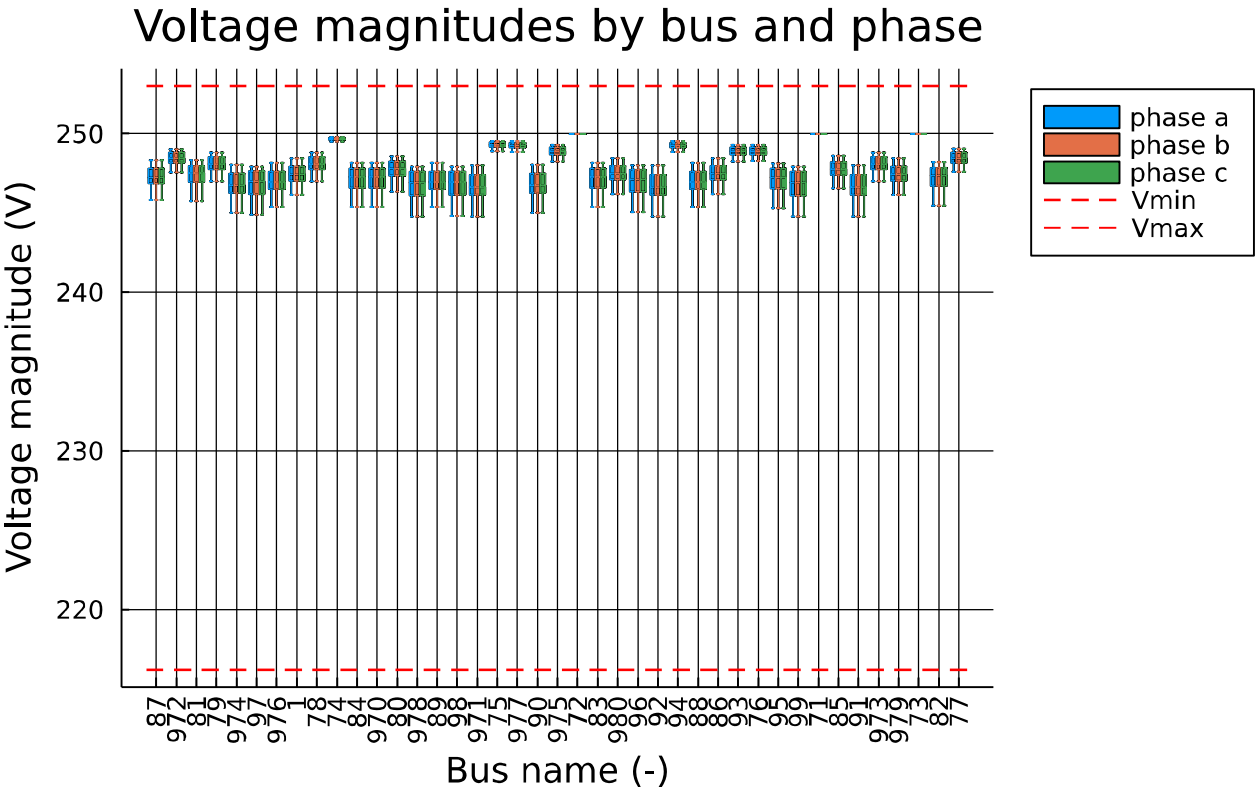
C.4 Network H


Power flow through substation bus
Active power (kW)
9
8
7
6
5
4
phase 1
phase 2
phase 3
5
10
15
20
Reactive power (kvar)
3.5
3.0
2.5
2.0
5
10
15
20
Time (hour)

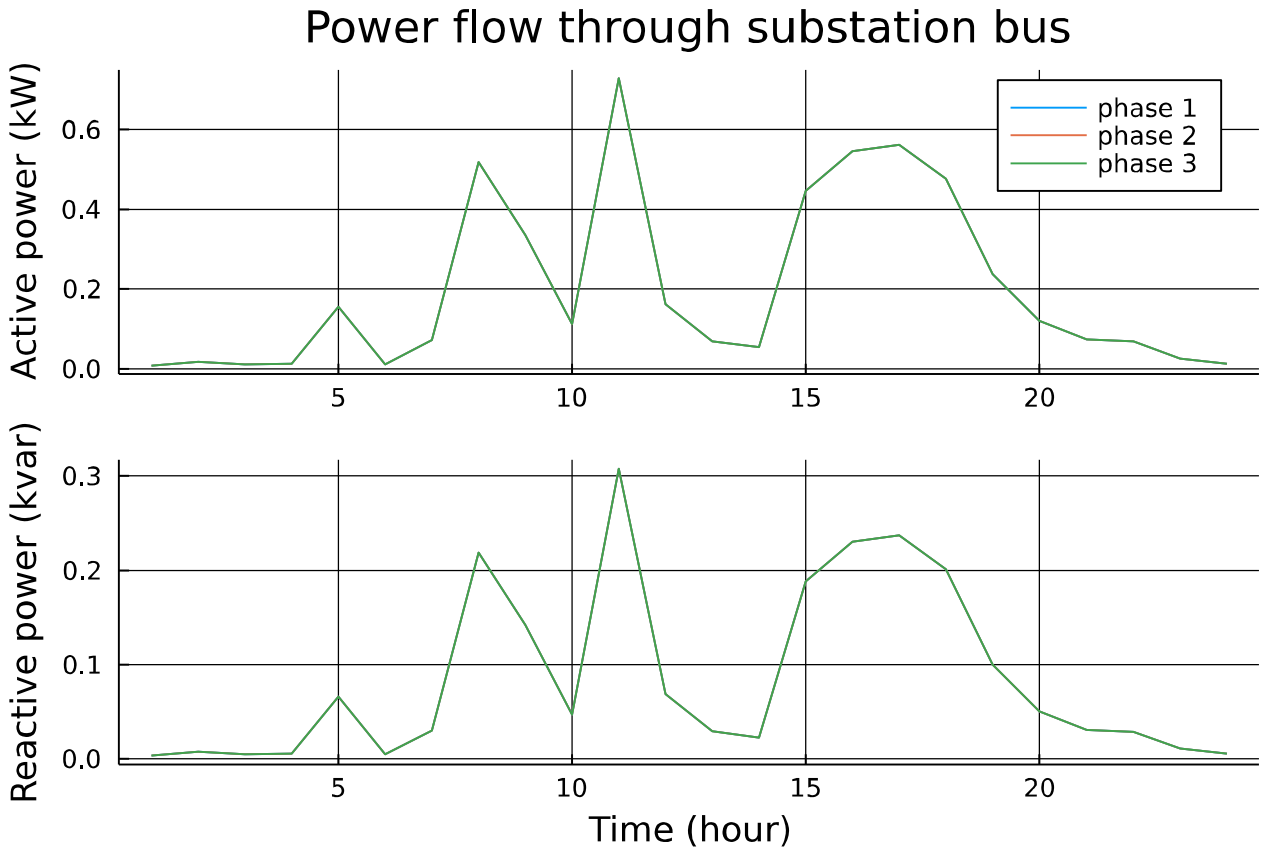

Voltage magnitudes by bus and phase
Voltage magnitude (V)
250
240
230
220
phase a
phase b
phase c
Vmin
Vmax
24
29
12
20
06
25
23
22
11
13
27
0978
15
0969
0976
28
16
14
21
0975
0973
09
0972
0968
0971
08
26
17
19
10
0966
0970
0977
18
0974
0979
Bus name (-)

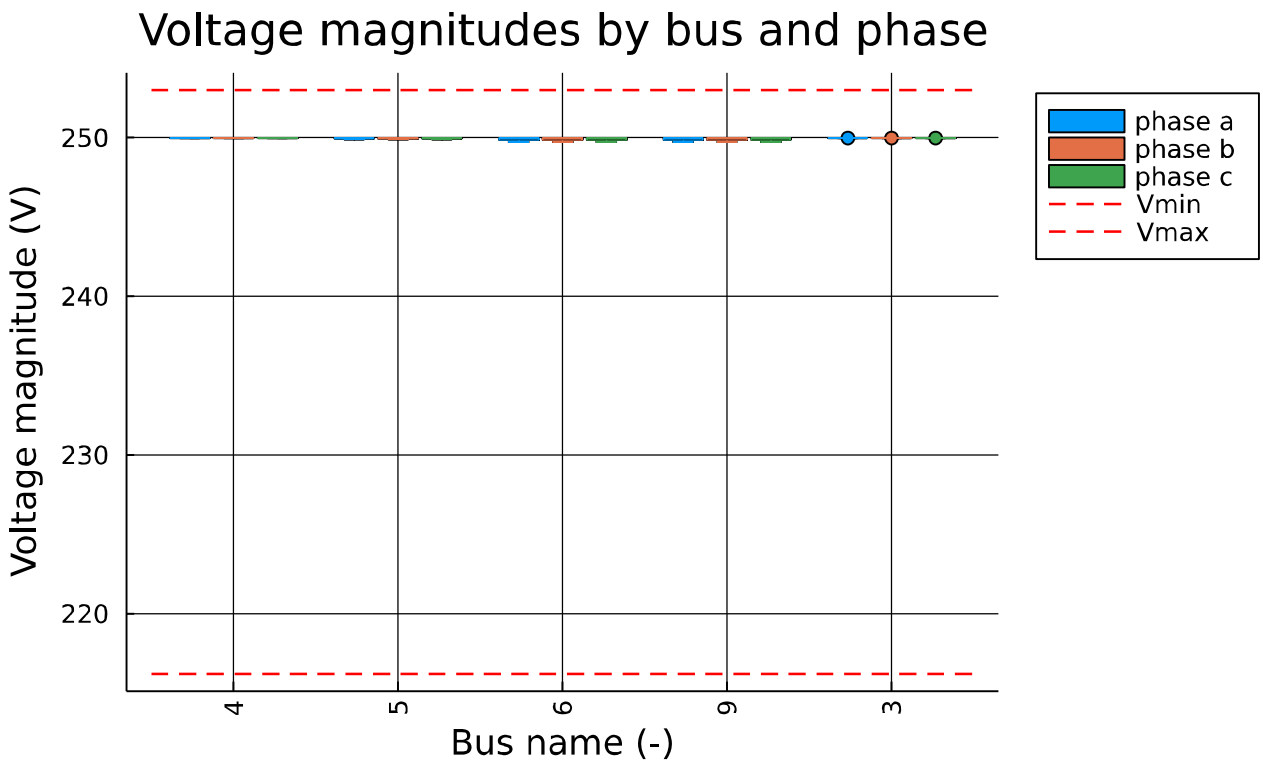
C.5 Network I


Power flow through substation bus
Active power (kW)
14
12
10
8
6
5
10
15
20
phase 1
phase 2
phase 3
Reactive power (kvar)
6
5
4
3
5
10
15
20
Time (hour)

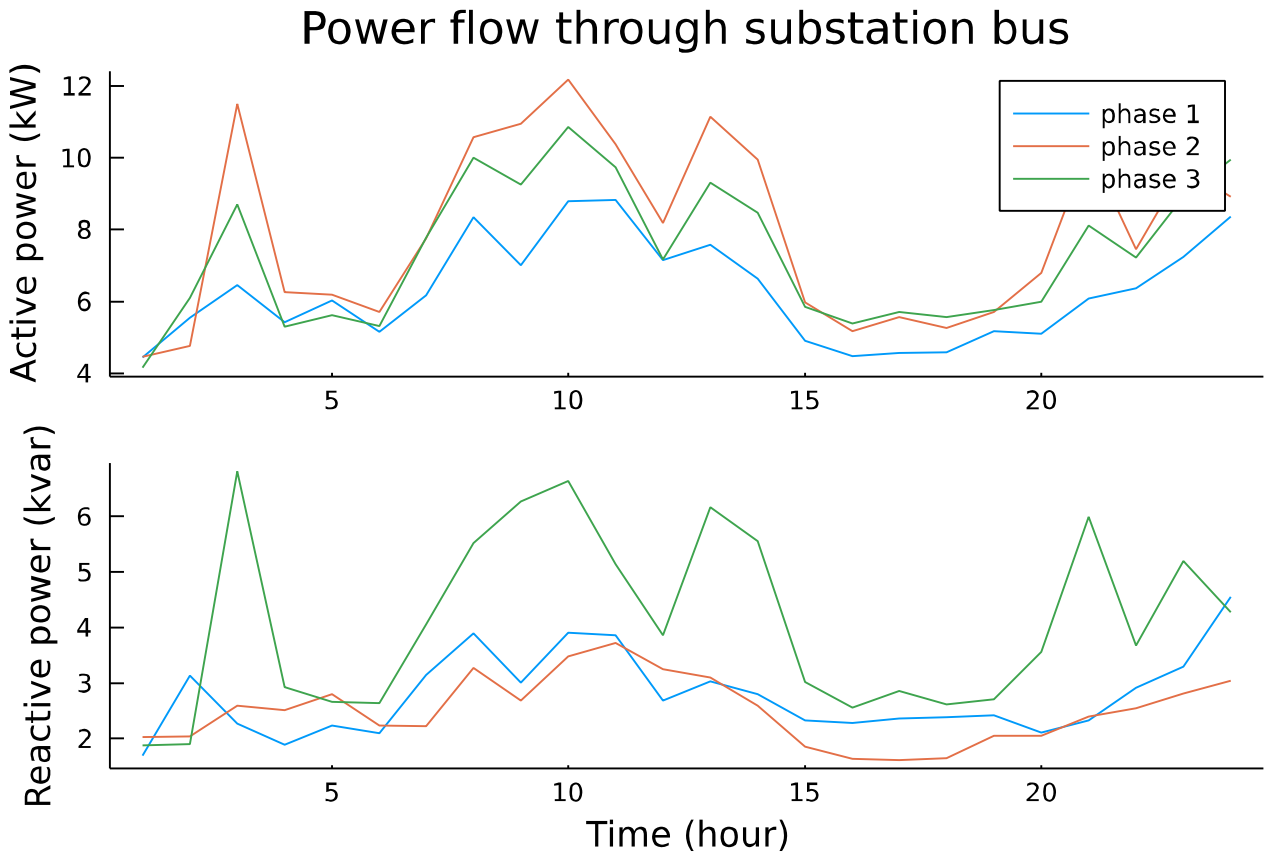

Voltage magnitudes by bus and phase
Voltage magnitude (V)
250
240
230
220
phase a
phase b
phase c
Vmin
Vmax
Bus name (-)

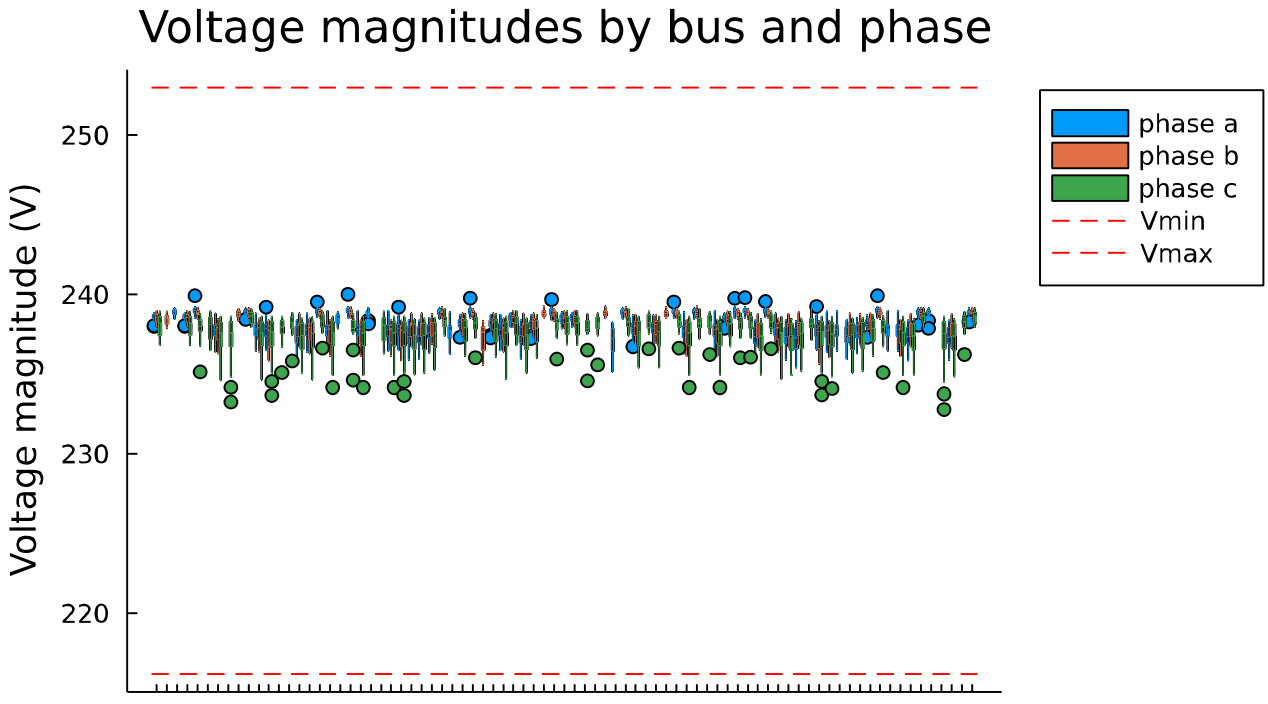
C.6 Network K


Power flow through substation bus
Active power (kW)
0.0
0.2
0.4
0.6
5
10
15
20
phase 1
phase 2
phase 3
Reactive power (kvar)
0.0
0.1
0.2
0.3
5
10
15
20
Time (hour)

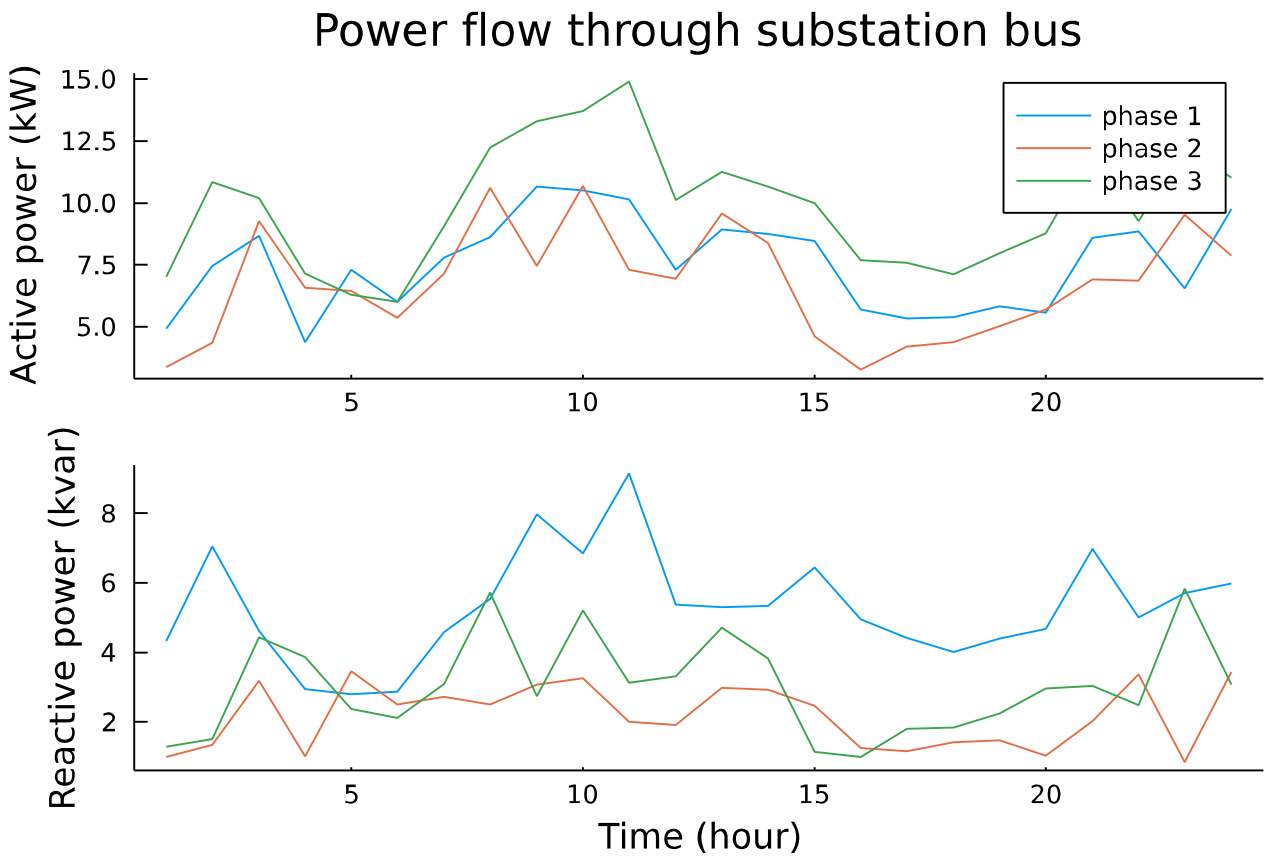

Voltage magnitudes by bus and phase
Voltage magnitude (V)
220
230
240
250
4
5
6
9
3
Bus name (-)
phase a
phase b
phase c
Vmin
Vmax

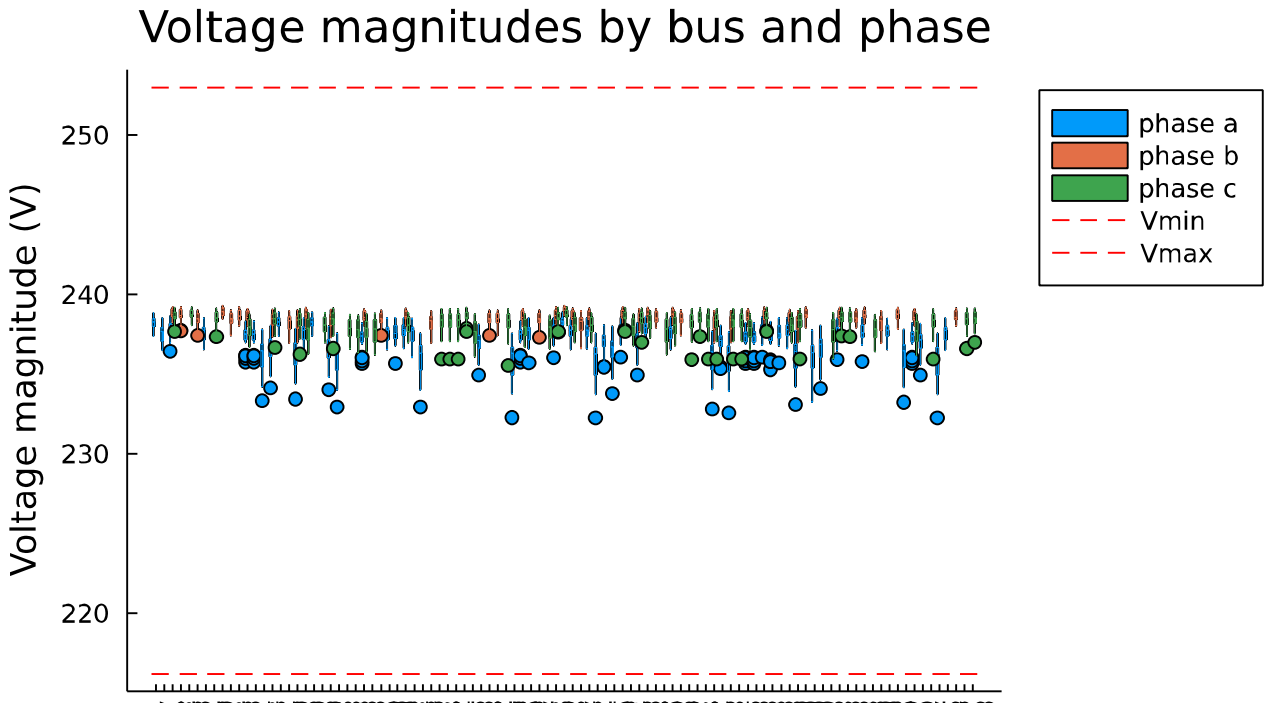
C.7 Network M


Power flow through substation bus
Active power (kW)
12
10
8
6
4
5
10
15
20
phase 1
phase 2
phase 3
Reactive power (kvar)
6
5
4
3
2
5
10
15
20
Time (hour)

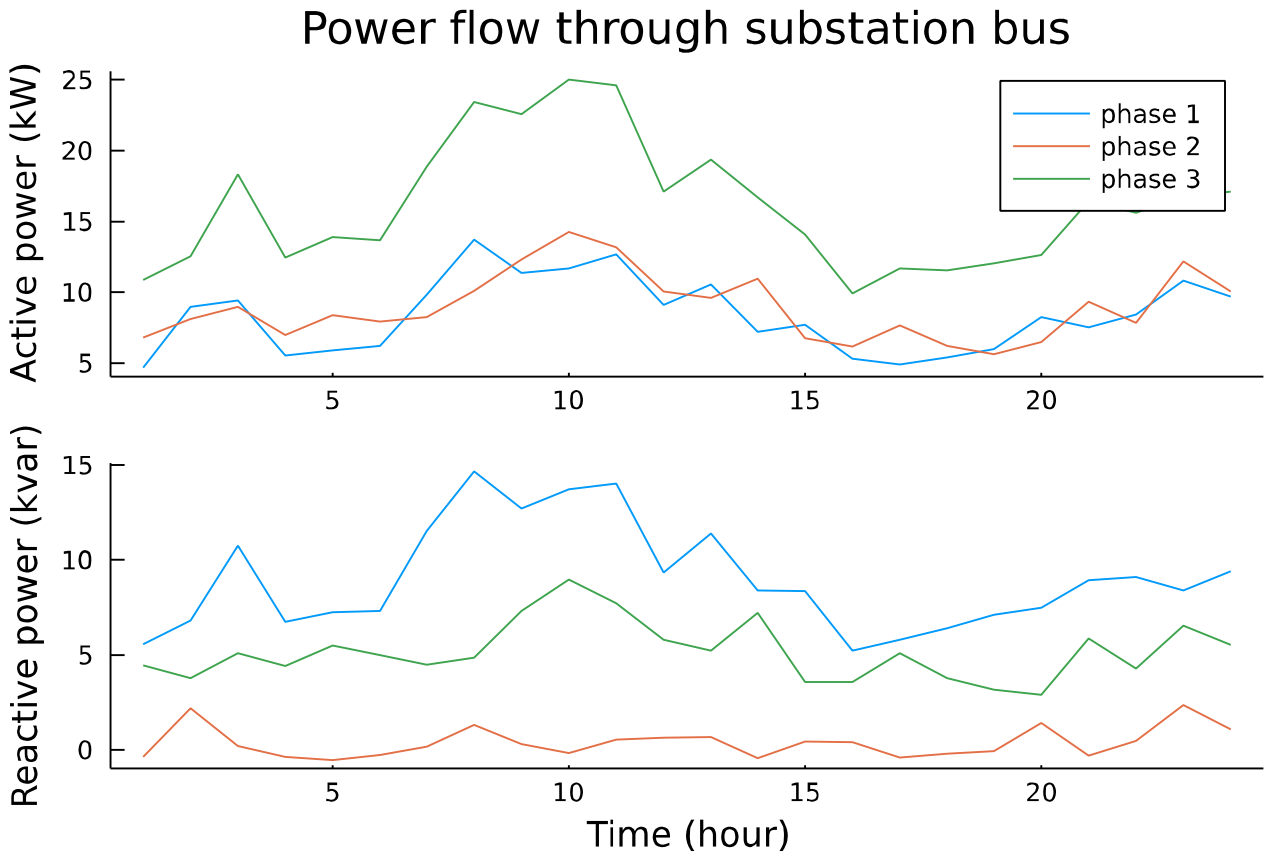

Voltage magnitudes by bus and phase
Voltage magnitude (V)
250
240
230
220
phase a
phase b
phase c
Vmin
Vmax

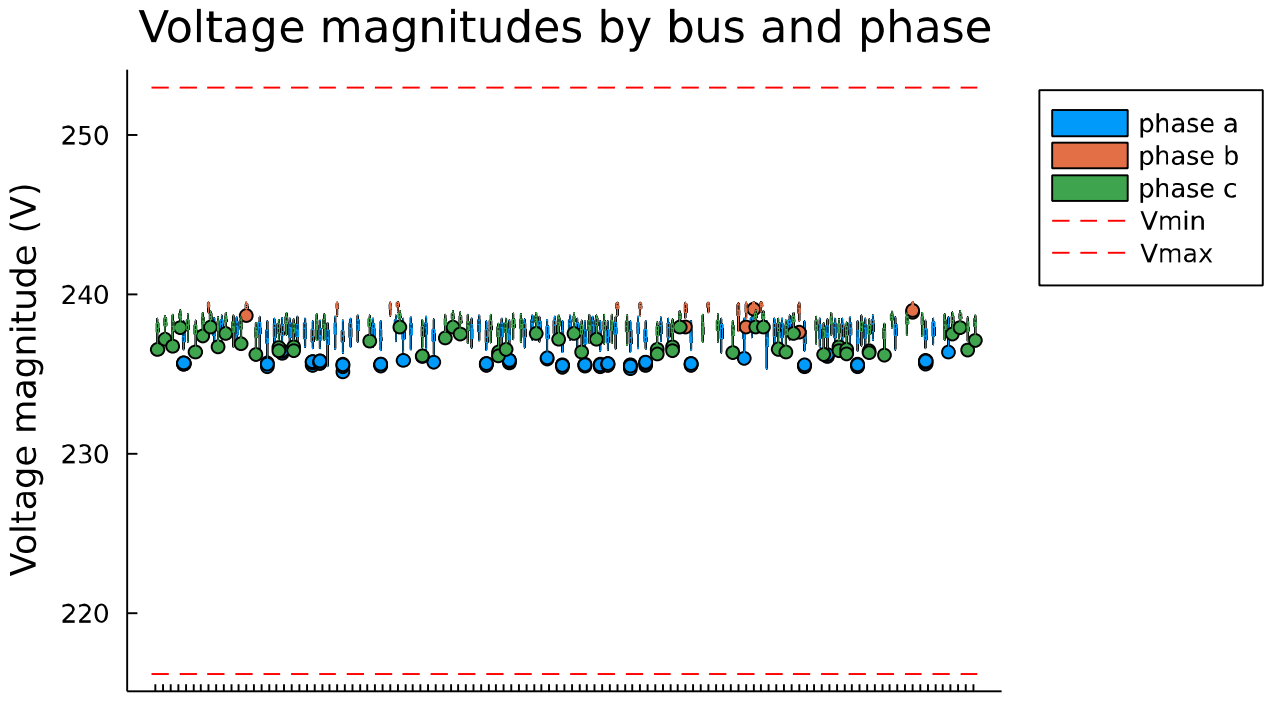
C.8 Network N


Power flow through substation bus
Active power (kW)
15.0
12.5
10.0
7.5
5.0
phase 1
phase 2
phase 3
5
10
15
20
Reactive power (kvar)
8
6
4
2
5
10
15
20
Time (hour)

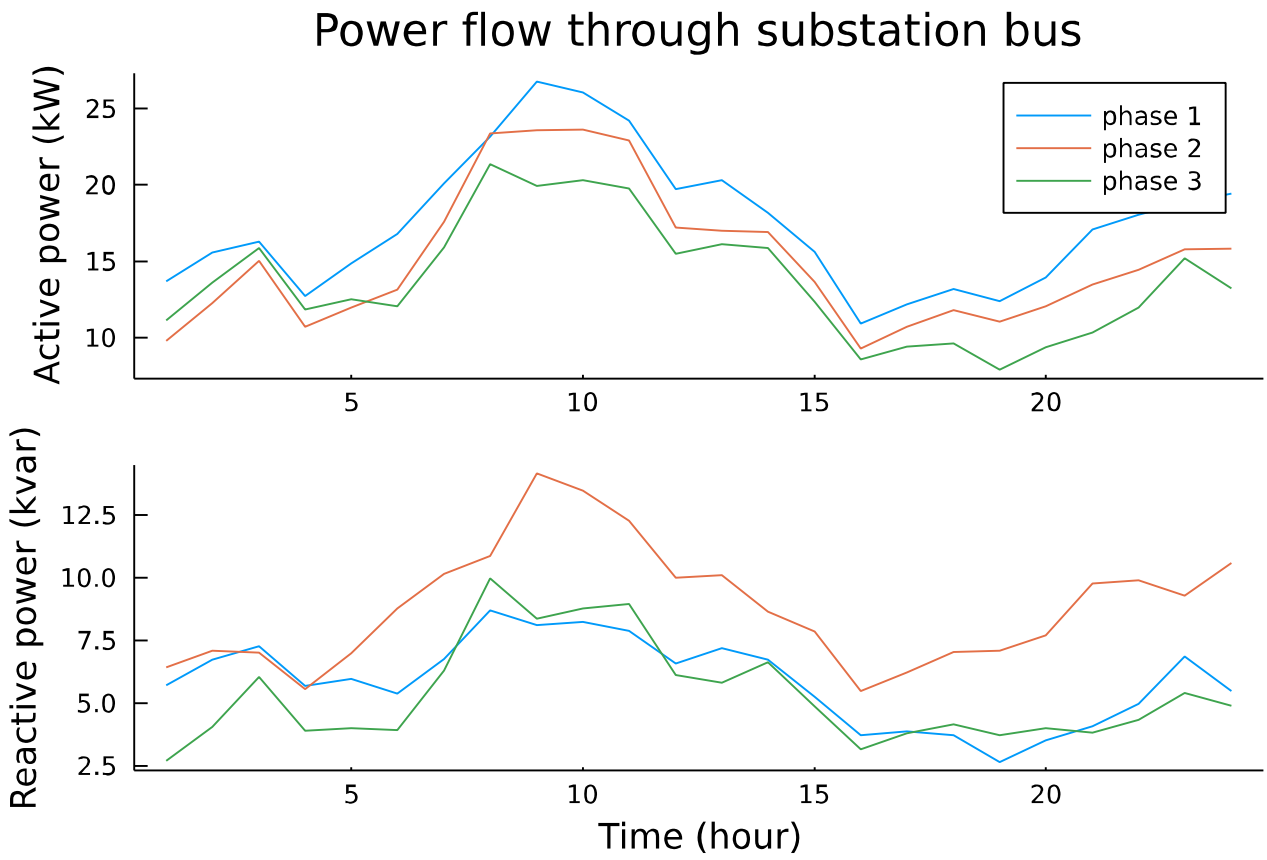

Voltage magnitudes by bus and phase
Voltage magnitude (V)
250
240
230
220
phase a
phase b
phase c
Vmin
Vmax

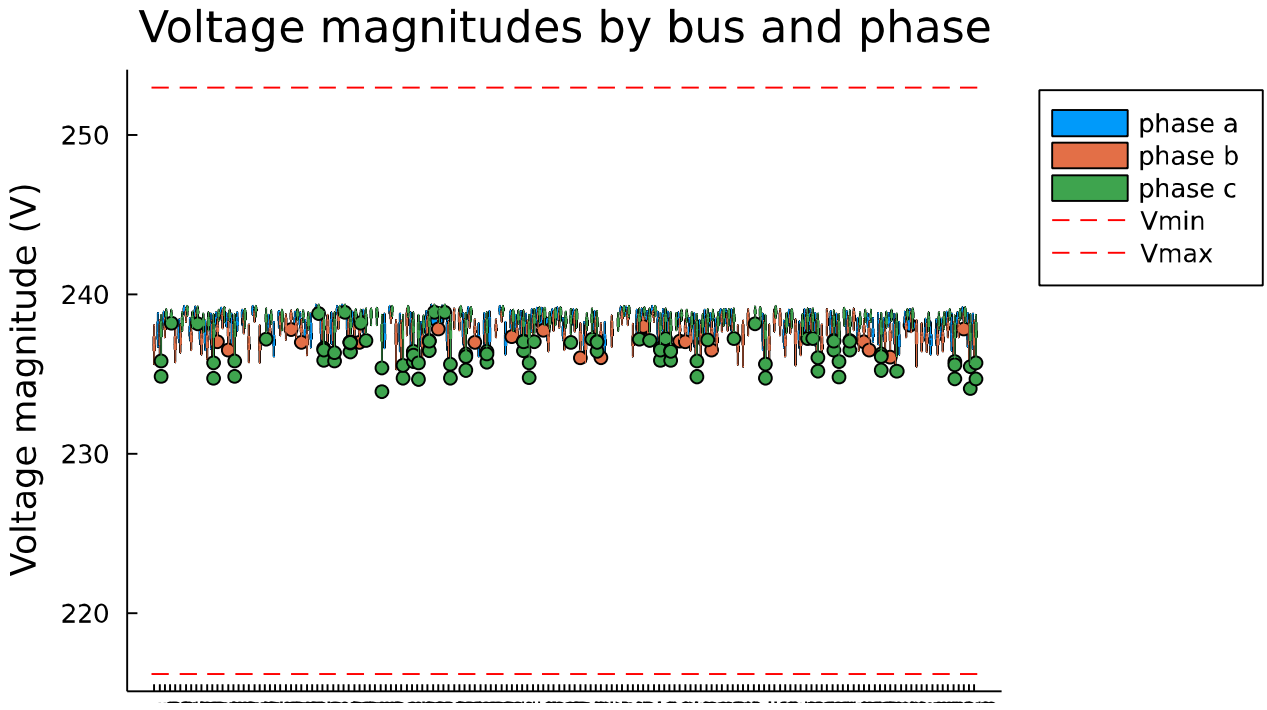
C.9 Network O


Power flow through substation bus
Active power (kW)
25
20
15
10
5
phase 1
phase 2
phase 3
5
10
15
20
Reactive power (kvar)
15
10
5
0
5
10
15
20
Time (hour)

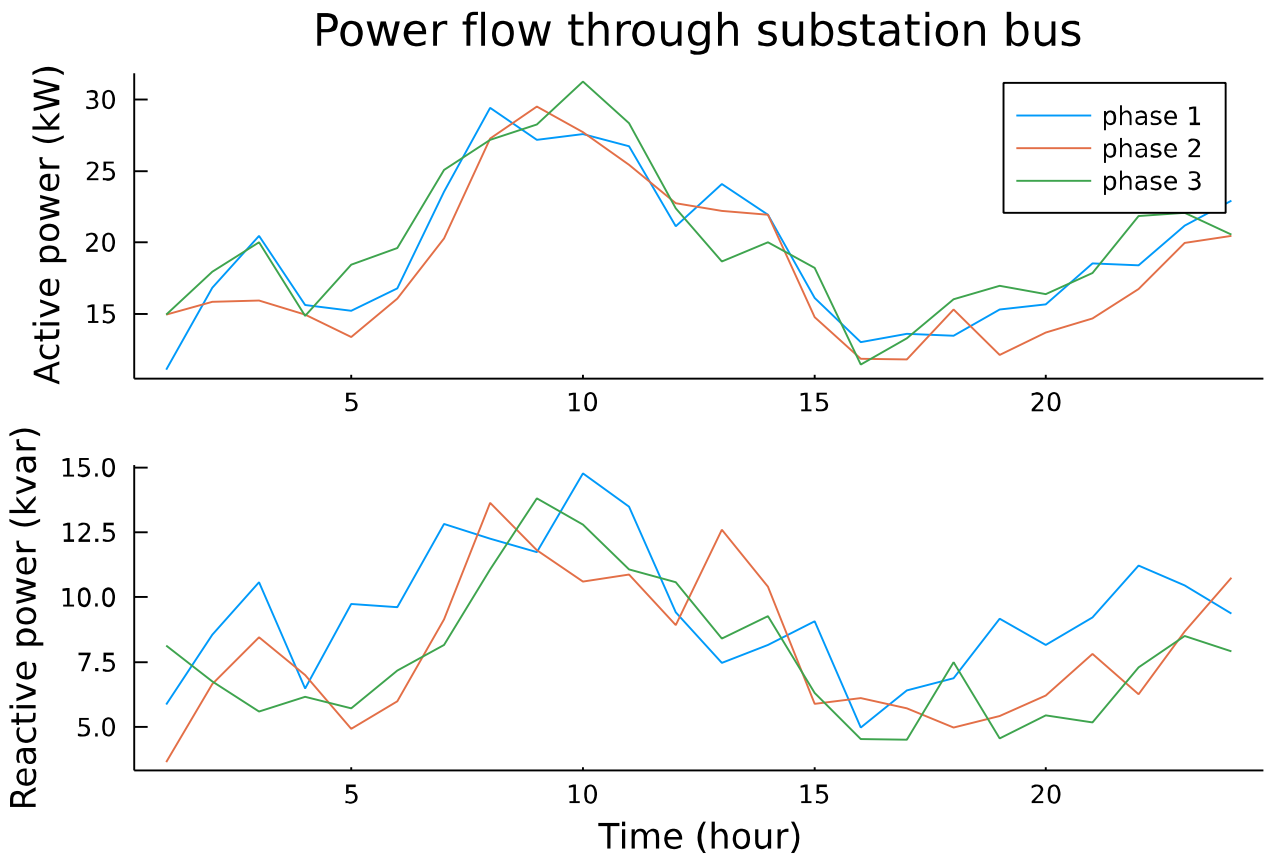

Voltage magnitudes by bus and phase
Voltage magnitude (V)
250
240
230
220
phase a
phase b
phase c
Vmin
Vmax

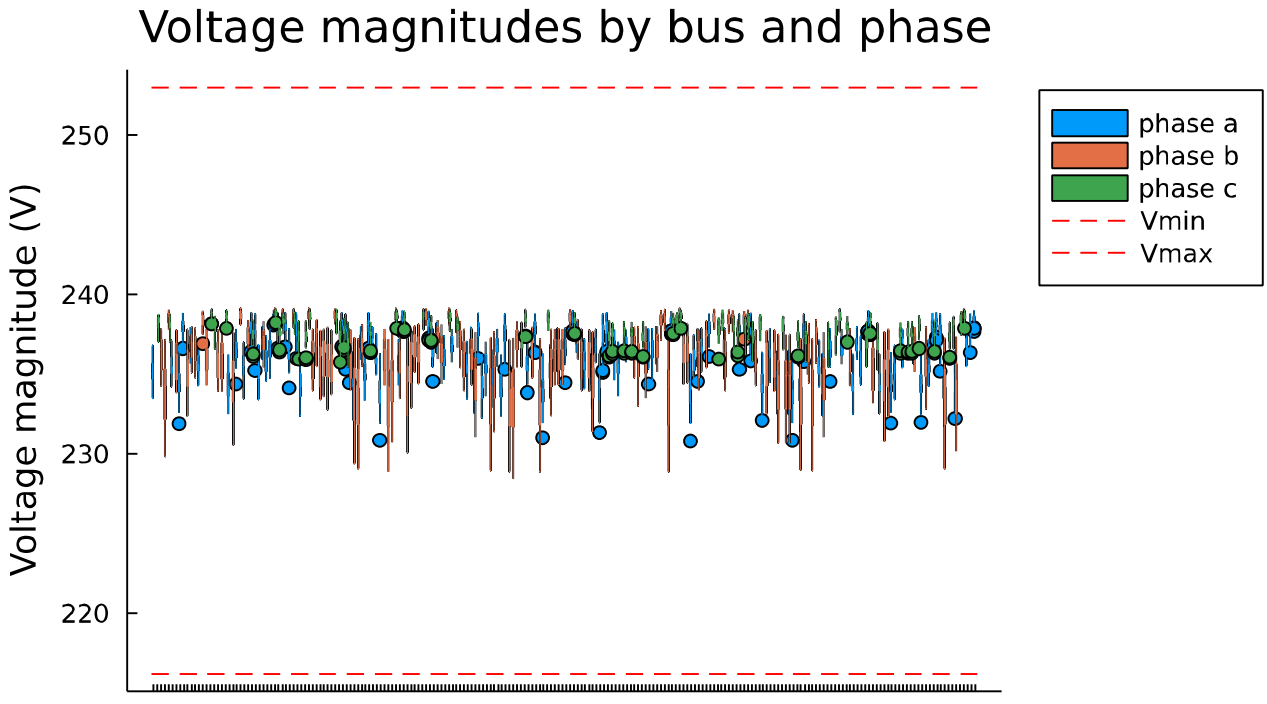
C.10 Network P



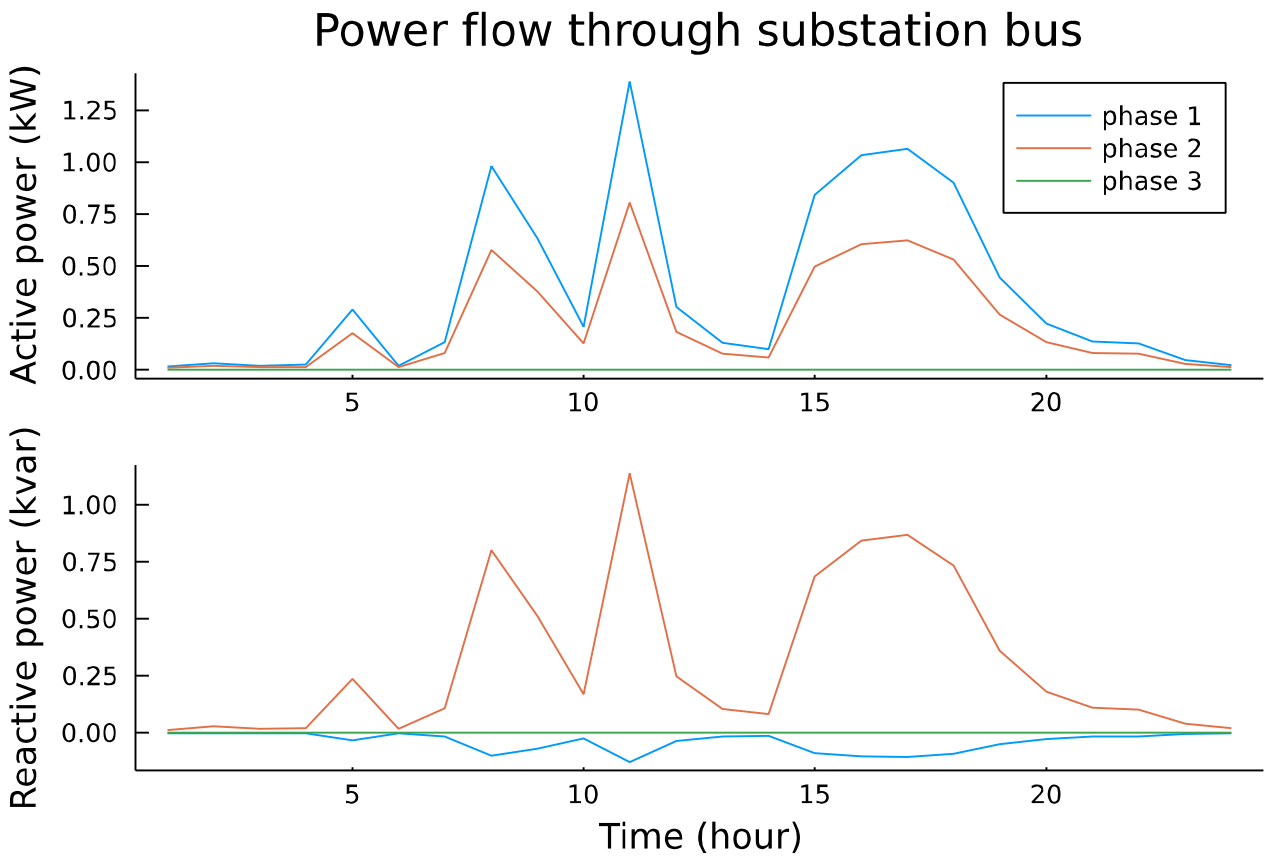


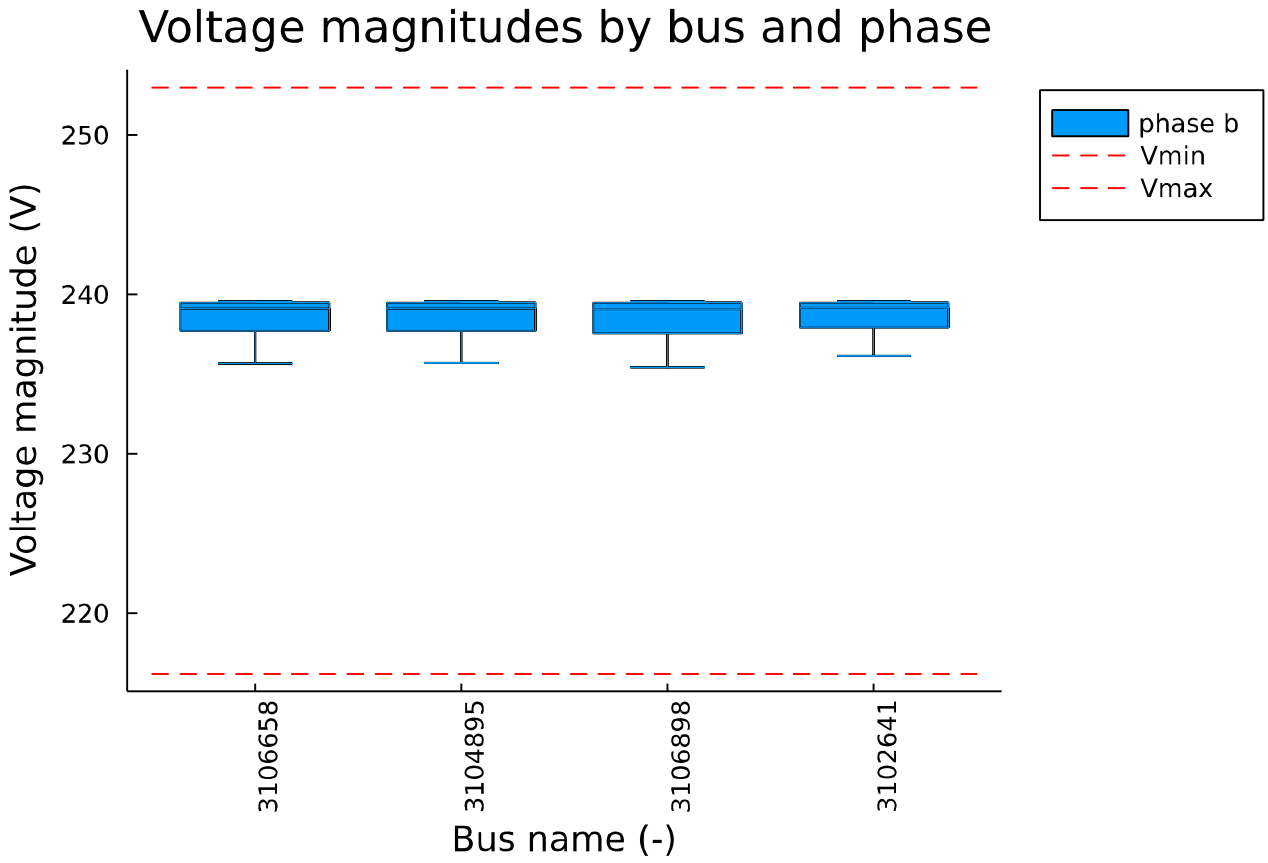
C.11 Network R



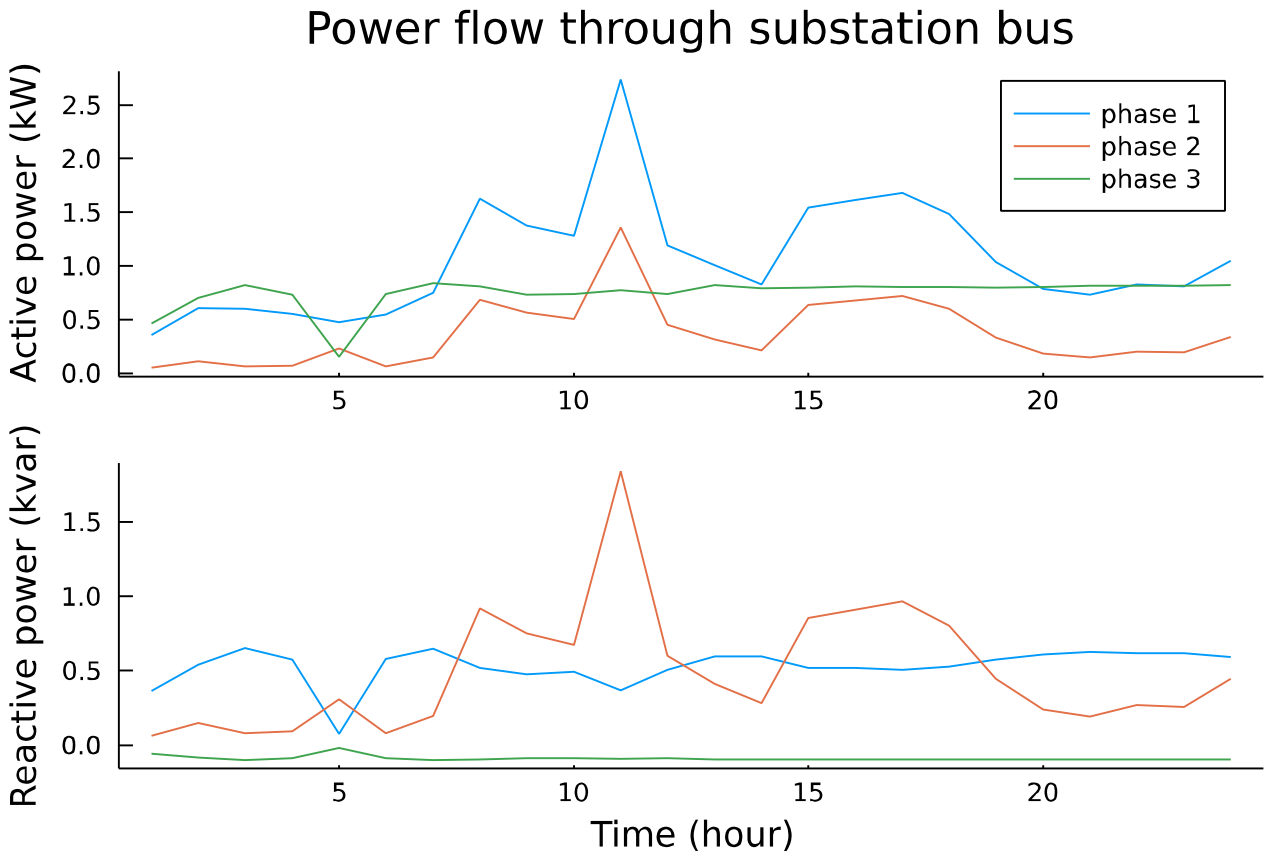


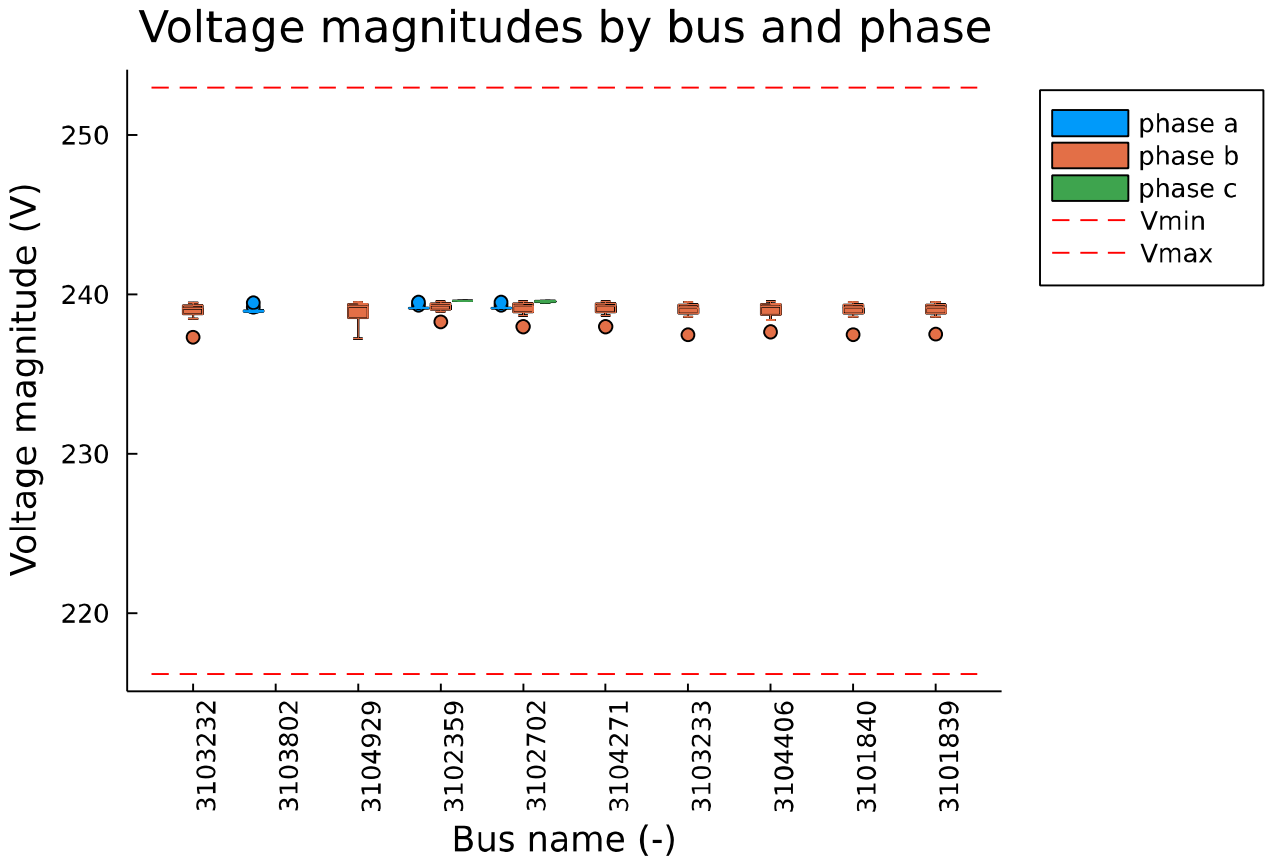
C.12 Network T



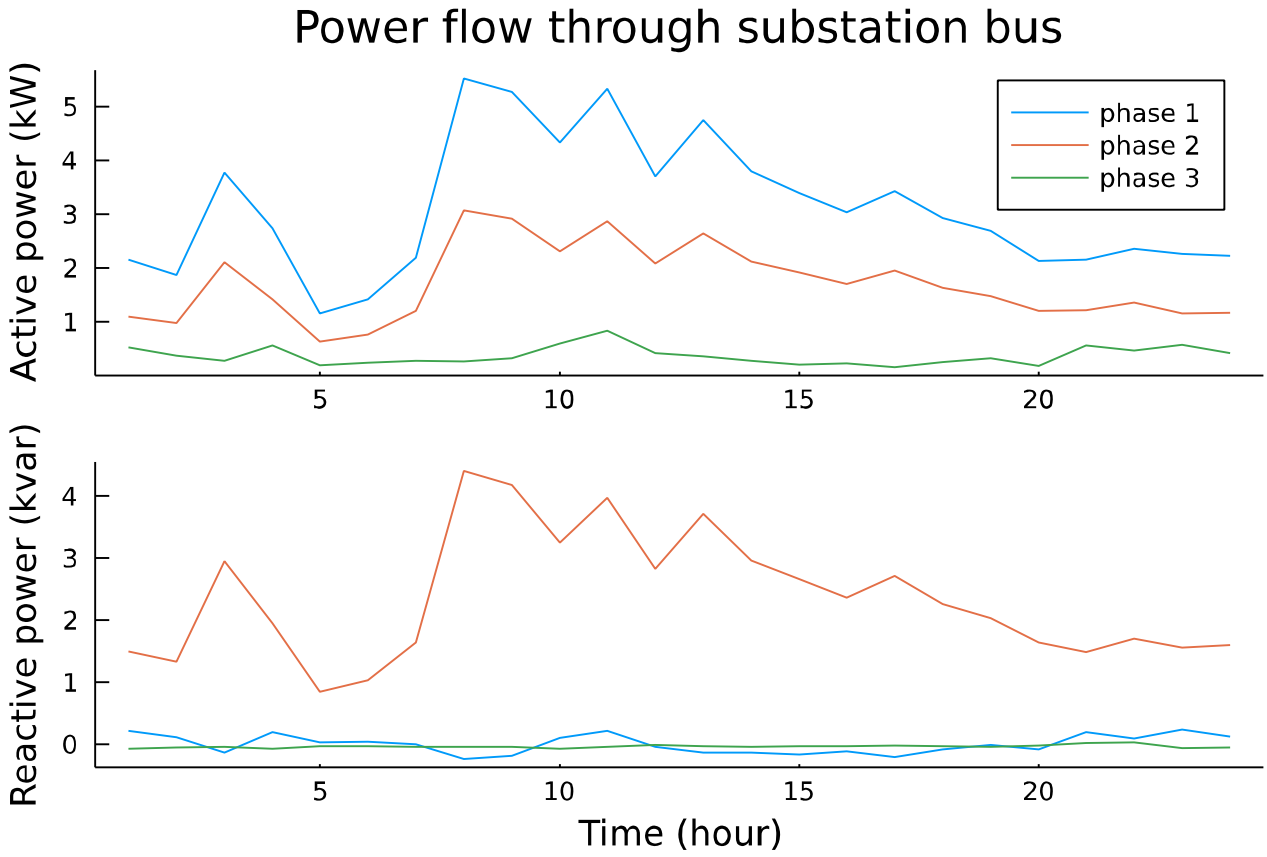


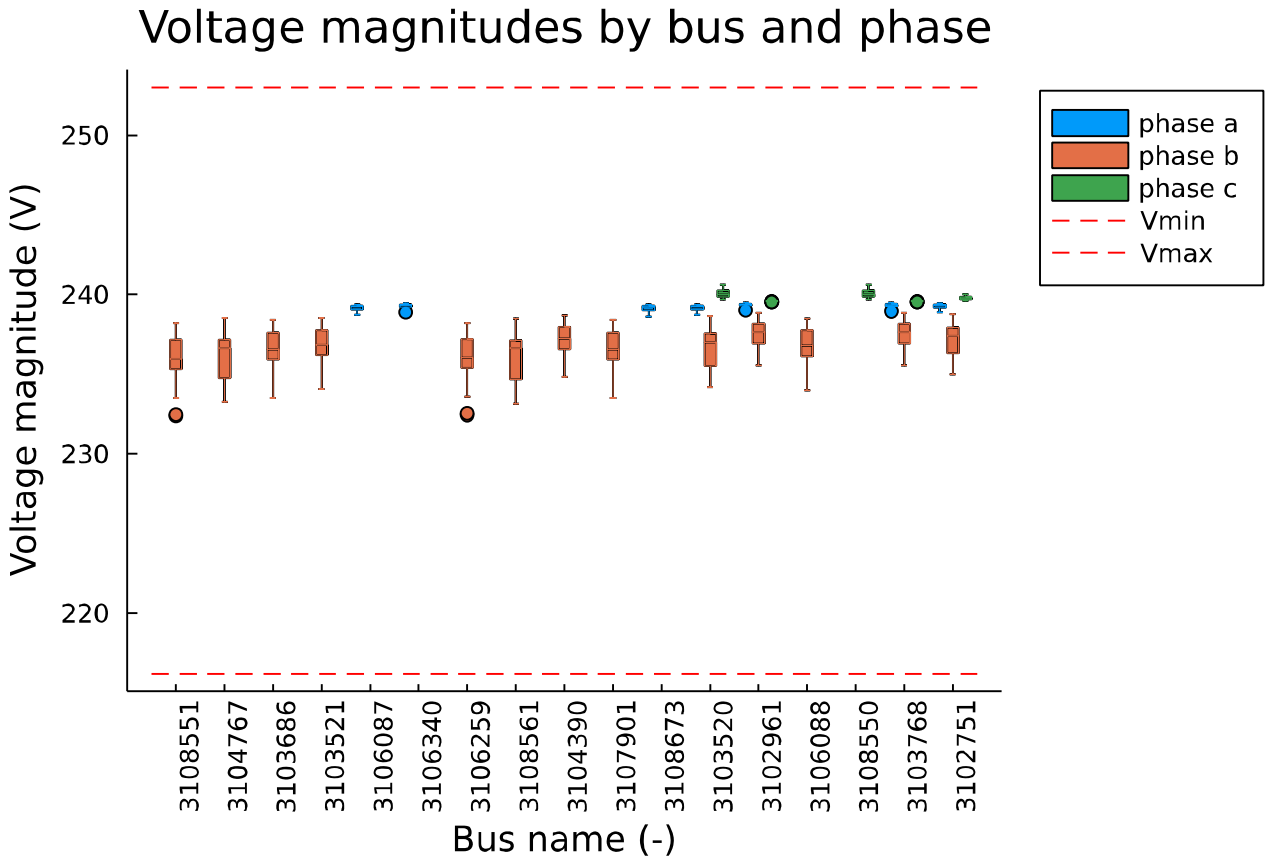
C.13 Network U



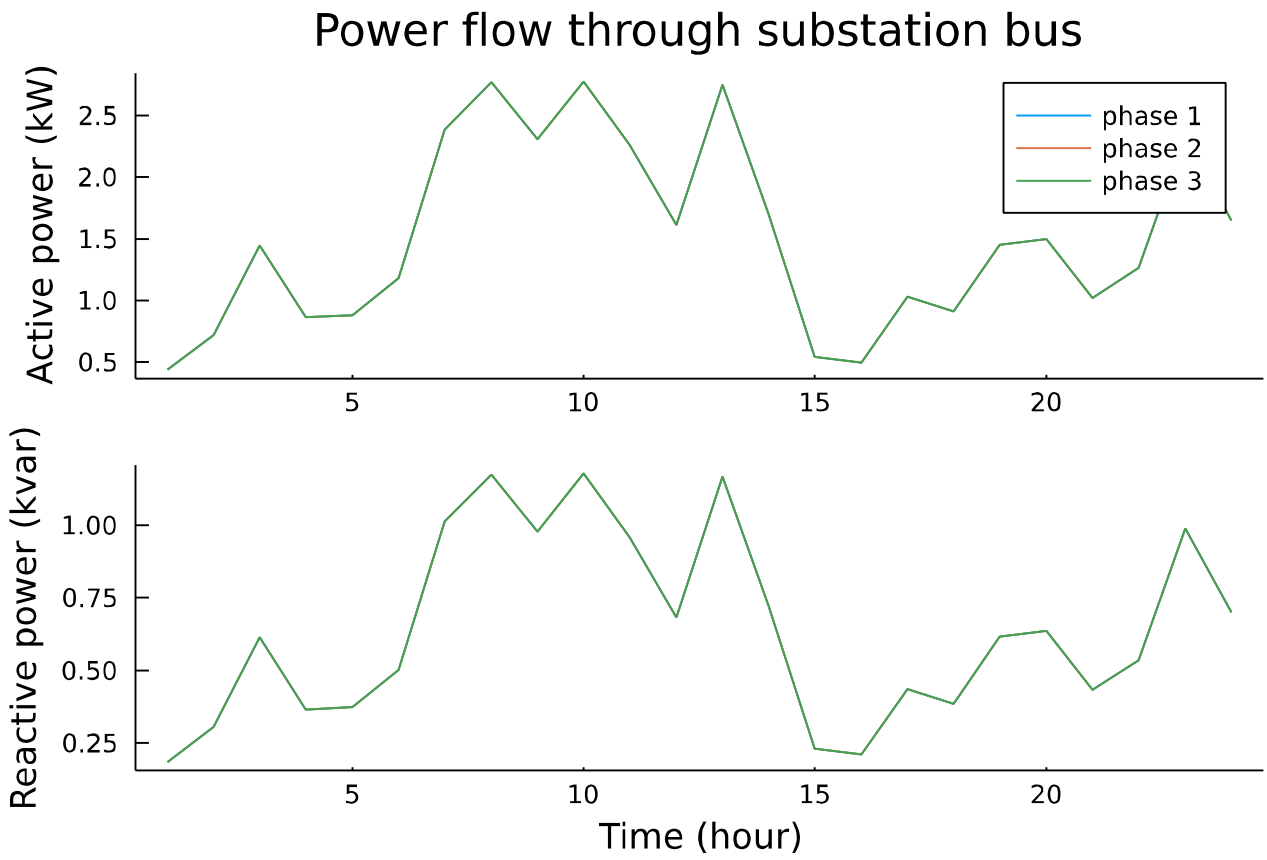


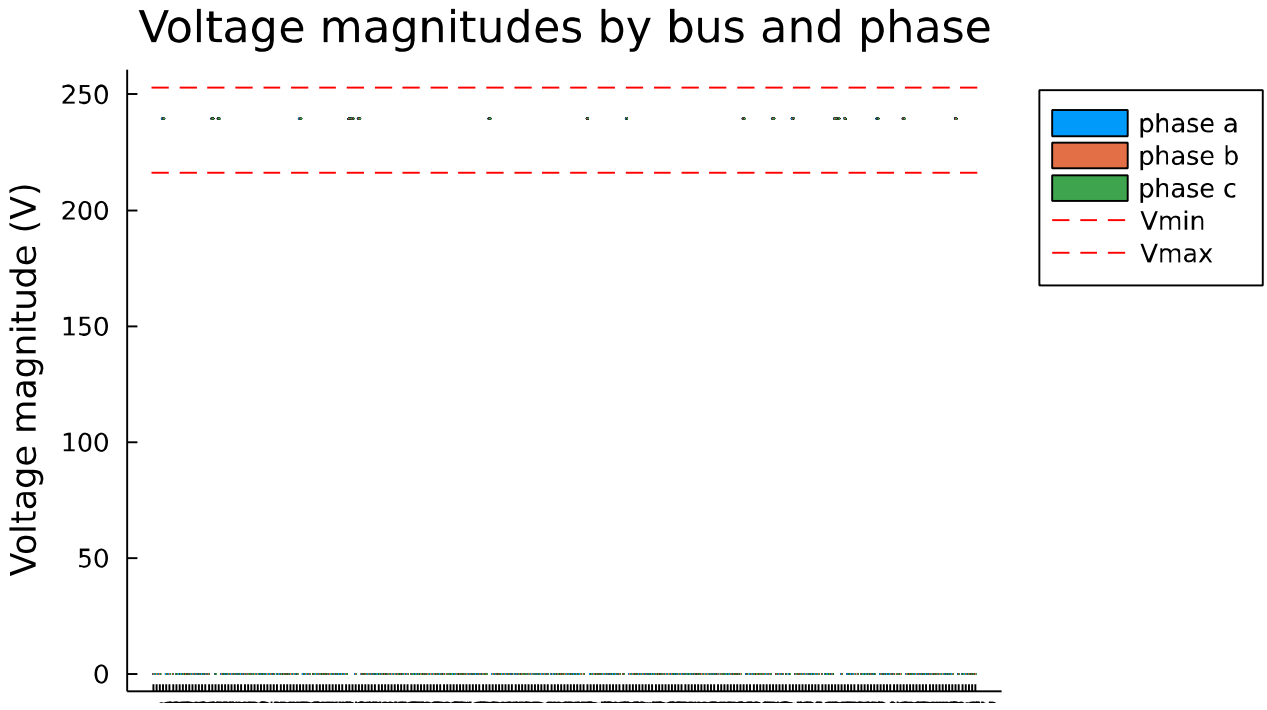
C.14 Network V





C.15 Network W





Appendix D Clustering with Updated Ausgrid load counts

Following the internal release of the final report and representative network models, Ausgrid identified an issue with missing load data on many of their nodes and raised valid concerns that this would skew any resulting network simulations outputs.

The issue was traced back to an error in the custom-built parser that was written to ingest the non-standard Ausgrid network data Excel format. Once corrected, the average number of loads per network increase from 10 to 20, over approximately 45,000 Ausgrid LV networks. As this appeared to be a significant change, it was decided to re-cluster the networks and compare to the published clusters to determine sensitivity to this input feature change. The cluster input features n_Load and num_customers were corrected, and the final k-medoids clustering was re-run with identical settings and random seed, i.e. the only changes were the two features dependent on load count.

Fortunately, the resulting clusters and medoids did not differ significantly from the previously identified results. The main quantitative metric, the Davies-Bouldin score increased (worsened) from 2.3 (the red 'final' dot in Figure 65) to 2.6, indicating that the density-vs-separation of clusters in the latter got worse. These numbers are not strictly comparable as the input data differed slightly but is a reasonable approximation when considered alongside the qualitative results.

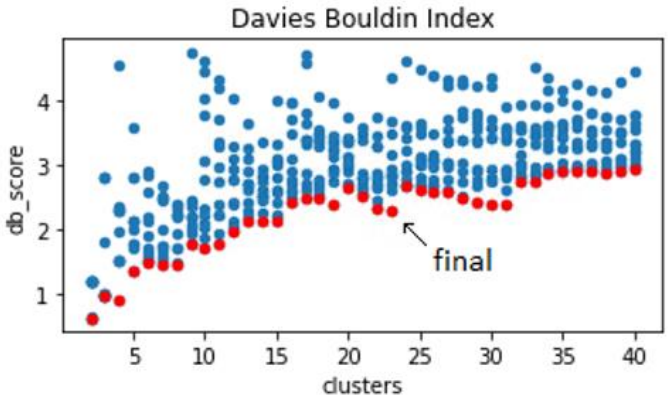


Figure 65 – final clustering DB-scores, before Ausgrid load-count correction. DB-score after correction increased (got worse) from 2.3 to 2.6.

Qualitatively, the network renders of the medoids before and after the feature correction were compared by domain experts, who suggested that the small changes in cluster medoids probably did not warrant switching to the newer set. The two sets of medoid renders are shown in Figure 66 and Figure 67, noting that the latter figure does not yet show the missing Ausgrid loads (though they will be included in the released network models).

The main conclusions from the 23 representative medoids were:

- 6 network medoids did not change

- An additional 11 clusters were relatively strongly consistent between the two clusterings (in the sense of having a large proportion of members in common)
- The size of the clusters did not appear to have a large change (although one initial cluster appeared to split, with one part remaining the split joining another initial cluster in the formation of the updated clusters).
- There was no significant change in the DNSP balance between clusters
    - There were still 9 clusters represented by an Ausgrid medoids,
    - TasNetworks decreased from 11 to 10 medoids
    - Essential remained at 1 cluster
    - SAPN remained at 2 medoids (though both very small clusters sizes)
    - Endeavour gained a medoid
- The general spread of network topologies was visually very similar – certainly more so than the differences between different values of k explored originally

This *lack* of significant changes to the clusters following an apparently significant change to the input data shows that the clustering methodology is actually quite stable and robust

Because there was not significant change to the clusters in these results it was decided to keep the existing set of medoids as the final released network models. The Ausgrid medoids had the missing loads added though, to allow valid simulation scenarios to be run on them, as presented in Sections 5.7, 6.1, 6.3 and Appendix C .

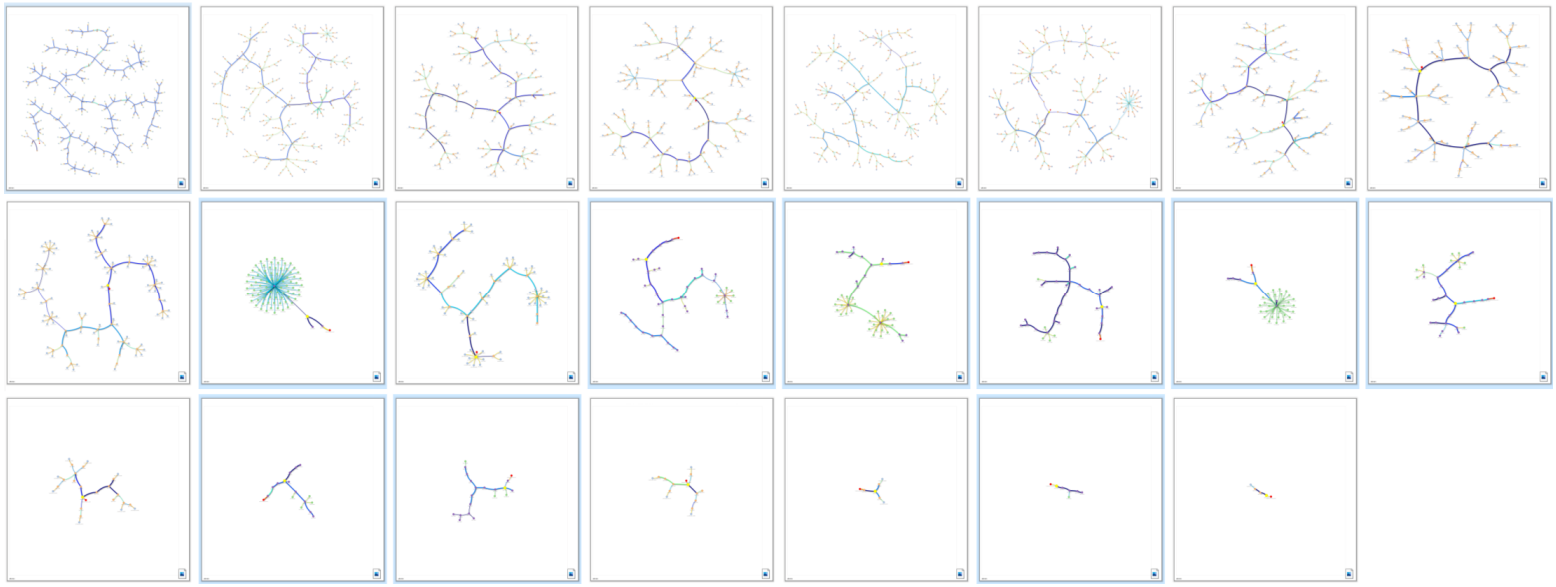


Figure 66 - Final cluster medoids, *before* Ausgrid load-count correction. Ausgrid networks are highlighted in light blue.

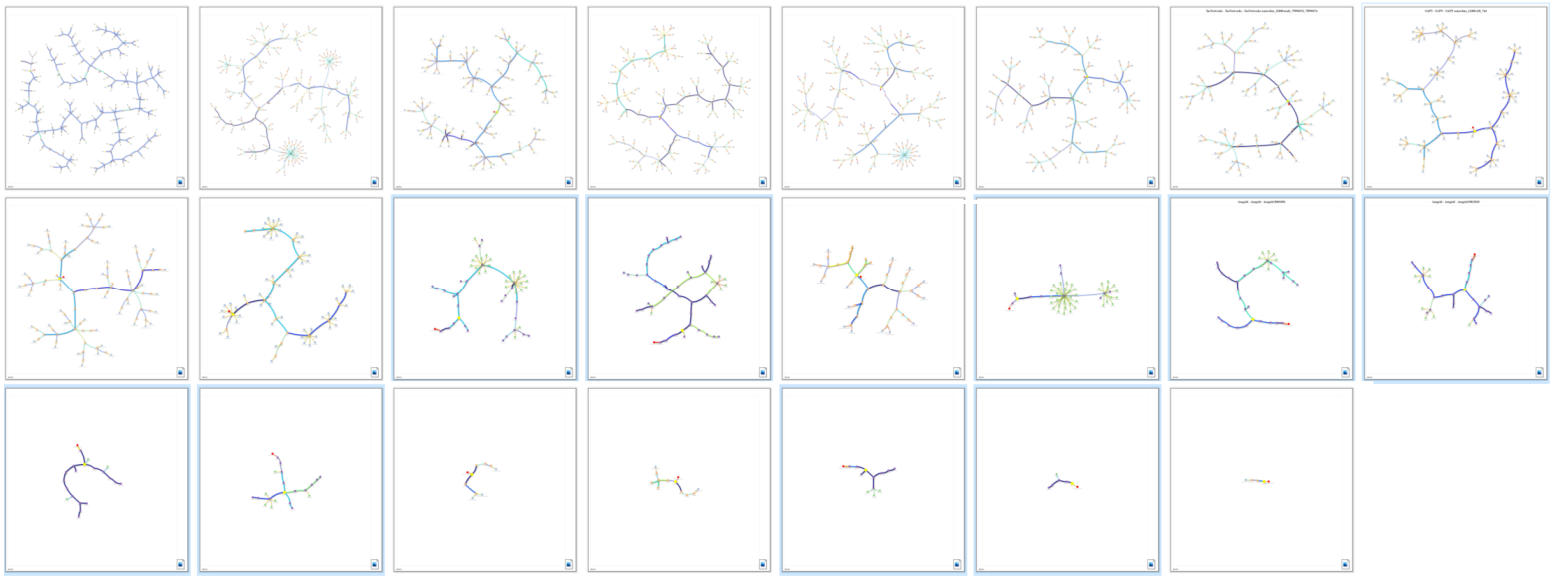


Figure 67: Cluster medoids, *after* Ausgrid load-count correction. Ausgrid networks are highlighted in light blue. Missing loads in Ausgrid networks are omitted from these renderings for ease of comparison. Clusters were calculated using the correct Ausgrid load-counts.

As Australia's national science agency and innovation catalyst, CSIRO is solving the greatest challenges through innovative science and technology.

CSIRO. Unlocking a better future for everyone.

Contact us
1300 363 400
+61 3 9545 2176
csiroenquiries@csiro.au
www.csiro.au

For further information
Energy Systems
Gavin Cross
+61 2 4960 6262
Gavin.Cross@csiro.au
csiro.au/energy