Estimating Realistic Hybrid-Synthetic Linear Power Flow Transmission Models

Ryan Sparks, Mikhail Chester, Nathan Johnson

ASU-METIS-25-TRS-002

July 2025

TECHNICAL REPORT

ESTIMATING REALISTIC HYBRID-SYNTHETIC LINEAR POWER FLOW TRANSMISSION MODELS

Ryan Sparks, Mikhail Chester, Nathan Johnson

Metis Center for Infrastructure and Sustainable Engineering
School of Sustainable Engineering and the Built Environment

Laboratory for Energy And Power Solutions (LEAPS)
The Polytechnic School

Ira Fulton Schools of Engineering
Arizona State University

July 2025

Report No. ASU-METIS-25-TRS-002

ABSTRACT

Power flow models of the bulk electric transmission system are widely used to reveal insights on system behavior. These insights are useful not only for grid planning but increasingly for looking at complex interactions between different infrastructures. However, many existing models are either inaccessible to researchers or lack geospatial significance due to the use of entirely synthetic data. The objective of this work is to develop a geospatially relevant linear power flow model of the bulk electric transmission system - first for a sample region and then generalized for the United States. The power flow model incorporates real data where it is available and fills in data gaps with synthetic network generation methods from literature. This approach produces geospatially representative models which are statistically similar to synthetic models, and which provide realistic insight as to the behavior of the underlying infrastructure. This is particularly useful in understanding infrastructure behavior that may be tied to the electric power system. It can also provide a simplified yet realistic framework for transmission planning and interconnection, a well-known bottleneck for many energy projects throughout the United States. The resultant model creates a foundation for study of the electric power transmission system where geospatial relevance is important and reactive power planning can be neglected.

CONTENTS

1. INTRODUCTION

The study of infrastructure interdependencies and the wide-scale effects of the design of the electric transmission system is an increasingly critical field of study given the myriad disruptions these systems are encountering. Electricity systems are both complicated and complex (Arianos et al., 2025), and increasingly involve difficult judgement-based decision making that often goes unnoticed by consumers (Pandey et al., 2019). Electric transmission systems also experience difficulties with long-term planning and lock-in and are subject to the pressure and complexity of infrastructure development (Helmrich et al., 2023). Owing to their size, cost, security, and permanence, important decisions and research must be conducted without perfect insights into the functioning of these systems. Although power infrastructure is a well-studied and well-maintained system with excellent data fidelity, data availability is generally limited. Frequently the data is protected for security or privacy concerns, or simply because private enterprises choose not to provide it.

There are several existing efforts to broaden access to real electrical geospatial infrastructure and provide open-access models of realistic power systems (Parzen et al., 2023) (Hörsch et al., 2018). However, these models have some limitations and assumptions that affect transmission line behavior. Others have developed entirely fictitious electrical models with a high degree of intricacy (Birchfield et al., 2018; Birchfield, Xu, et al., 2017; Sadeghian & Wang, 2020). These models also have limitations, particularly in their geographical relevance and correlation with other infrastructures. To address this gap, we propose a statistical method following procedures outlined in (Sadeghian et al., 2018) for building realistic hybrid-synthetic linear power flow models for a transmission system with sufficient real data. This method outlines a process for constructing a hybrid-synthetic transmission network with basic network input data, and the implementation of a stochastic approach for estimating line characteristics, using Python network objects in PyPSA - an open-source, freely available software for power system modeling and power flow analysis (Brown et al., 2018).

2. AVAILABILITY OF GEOSPATIAL DATA

In an age of increasing data availability, real information is increasingly available in many locations concerning the structure and operation of the power grid (Borkowska & Pokonieczny, 2022; Haklay, 2010). Challenges exist in incorporating real data in synthetic network models, and several existing open infrastructure modeling efforts make high-level assumptions about the transmission system (Abdel-Khalek et al., 2025; Hörsch et al., 2018; Medjroubi et al., 2017; Wiese et al., 2019). For example, some of these models make implicit assumptions about the conductor type, bundle size, and tower construction, assigning either the same value to the entire system or assigning values based on voltage level (Parzen et al., 2023). This approach results in functional systems which may be suitable for study but lack statistical basis. Athari and Wang suggest there is no correlation between voltage level and distributed reactance (Ω/km), but there is a correlation with overall per-unit reactance (Athari & Wang, 2017). This correlation reveals a clear exponential distribution for total per unit reactance on each line based on voltage level. This statistical approach is used for modeling some synthetic grids and is suitable for augmenting real data for our hybrid-synthetic approach (Sadeghian & Wang, 2020).

DC power flow models are simpler than full AC models, but can still be used effectively in the study of transmission system behavior (Simpson-Porco, 2018; Trpovski & Hamacher, 2019; Villasana et al., 1985). Their computational efficiency allows for quick analysis of large-scale systems and numerous contingency scenarios, making them invaluable for a high-level overview of system performance (Stott et al., 2009). They are also useful in identifying potential bottlenecks and overloads during transmission planning studies, and are much friendlier in terms of data requirements. Nonlinear AC power flow studies are significantly more detailed and capture reactive power flow, voltage magnitudes, and stability issues, while their linear counterparts do not (Farrag et al., 2019). However, DC models still provide a reasonable approximation of real power flow and can effectively assess the impact of new lines, generation, loads, cascading failures, and congestion (Stott et al., 2009; Trpovski & Hamacher, 2019). This makes them a crucial tool for high-level study of the electric transmission system and its interdependence with other infrastructures. Linear power flow models with realistic significance can broadly contribute to the robustness and reliability of the electric transmission system.

3. HYBRID-SYNTHETIC MODEL

This work will implement a linear power flow transmission model built using widely available data, augmented by synthetic approaches to fill in data gaps. As a proof-of-concept, this power flow model will be constructed first for a sample region, then generalized for the continental United States. The realistic networks developed incorporate geospatially real data alongside synthetic values, producing a robust model with geospatial relevance. The input includes some basic data describing both the electrical network and population, and the output is a linear power flow model for the region. Geospatial network data are often available (such as those from the Department of Homeland Security (U.S. Department of Homeland Security, 2025)) and usually contain sufficient information for a geospatially relevant power flow model. Linear power flow models also require geospatially correlated load data (Landsman et al., 2024). Population data can also be used in lieu of load data when it is not available such as that available from the US 2020 Census (U. S. Census Bureau, 2020). These data sets at a minimum include geospatial layers for the following items:

- Transmission lines (with voltage attributes)
- Substations
- Power plants (with generation capacity attributes)
- Electricity consumption or total population
- Boundary file

A sample region is selected as an example of the network synthesis process, and the process can be conducted on areas larger or smaller than the sample region, as long as the data requirements are met. A subregion of WECC encompassing Arizona, New Mexico, and Southern Nevada (U.S. Energy Information Administration, 2020) is shown in Figure 1. Additional consideration should be given to the size of the sample region related to the use case of the network. Any representative model should include major sources of generation and should ensure that the load does not exceed generation. Load exceeding generation may be an indicator that the studied region is too small and should be expanded. Transmission lines which intersect with the regional boundary are included in the model, even if the ends of the lines may exceed the boundary of the region. Substations and power plants within the sample region or within a fixed

proximity to a transmission line are included in the modeling process. Total population data is available extracted for the sample region at the census tract level and any census tract that overlaps the sample region is selected. The boundary of the sample region is later used to assign substation service areas, which are created and then trimmed.

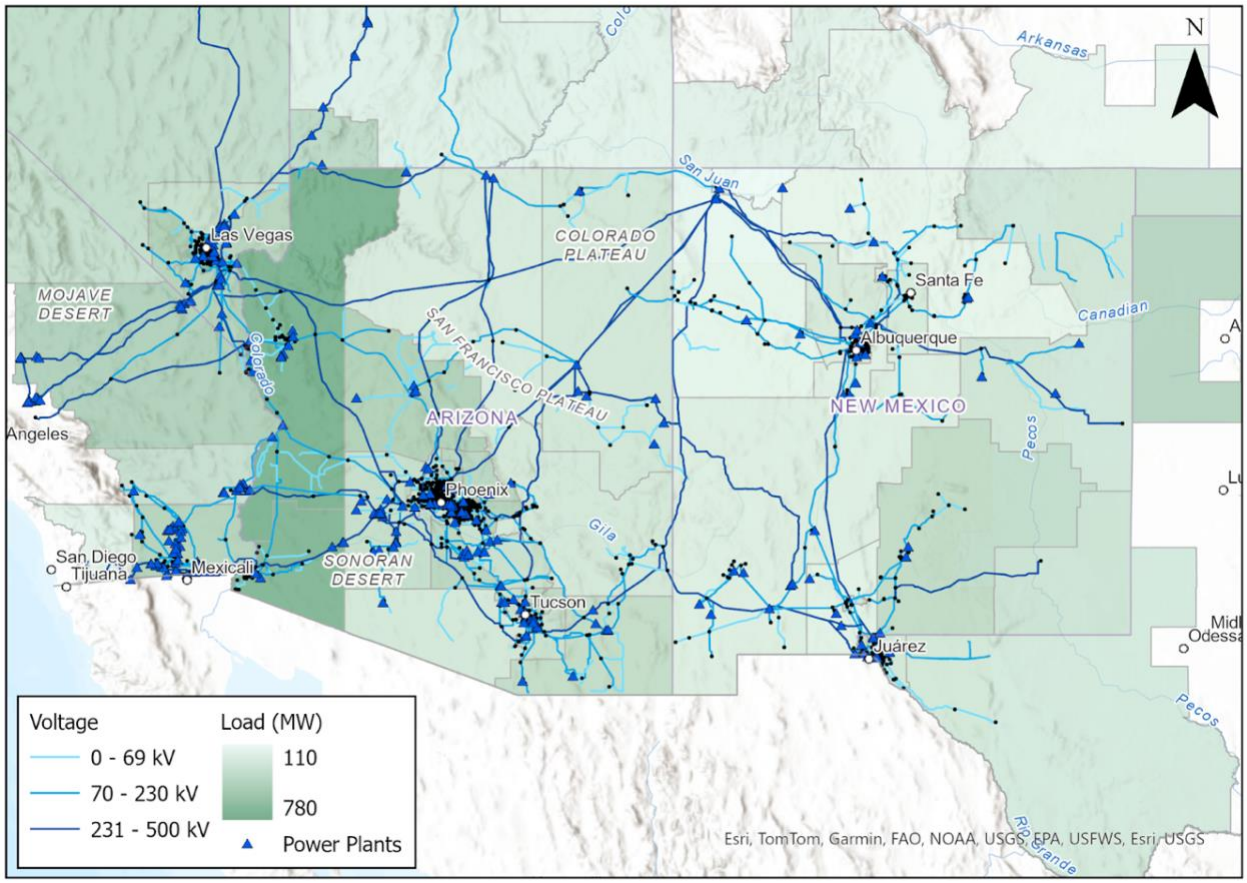

Figure 1: Input data for the WECC AZ-NM-SNV subregion

3.1. Network Construction

The construction of the DC power flow networks is conducted over six steps as shown in Figure 2. These steps happen sequentially to clean and process the GIS data into a working linear power flow model. Several assumptions are made which impact the performance of the system and are described for each step. Exceptions that arise during execution of the process are indicated to the user in case additional manual processing of the data is necessary.

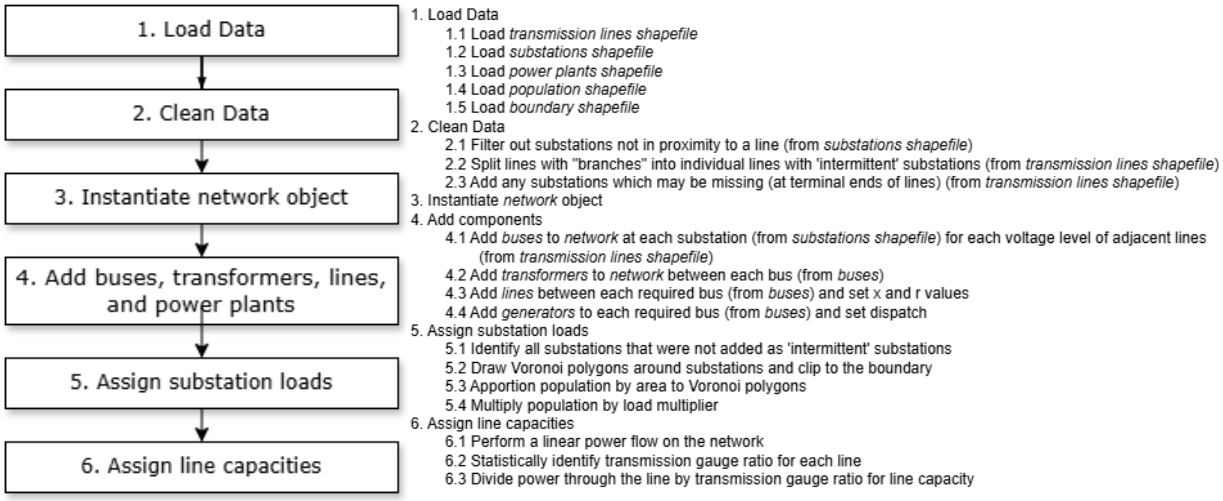

Figure 2: Network object construction process

Additionally, several assumptions are made through the model development process. These assumptions are to estimate network characteristics that are not readily available in public datasets, however where real data are available it should be used in lieu of these assumptions. The assumptions are summarized in Table 1. Additionally, synthetic approaches for statistically generating data such as line resistance and reactance, generator dispatch, and line capacities are used throughout the modeling process.

Table 1: Assumptions for model construction

Parameter	**Value**
Minimum substation distance from line	40 m
Minimum plant distance from substation	5 km
Transformer reactance	0.016 pu (Birchfield, Gegner, et al., 2017)
Transformer X/R ratio	80 (Birchfield, Gegner, et al., 2017)
Per unit MVA base	1 MVA (Brown et al., 2018)
Load distribution	NREL ResStock 2021 AMY2018 NREL ComStock 2024 AMY2018 (Landsman et al., 2024)

3.2. Data Cleaning

The subsequent steps in the network building process make several basic assumptions about the incoming dataset. Raw data may not strictly adhere to these assumptions and must be cleaned to maximize the functionality of the rest of the process and minimize loss of data. This data cleaning process works to find incongruities in the raw data that may easily be corrected. This includes issues such as distribution substations that have been misidentified as transmission substations, different lines which are combined into one layer, etc. Cleaned data should adhere to the following assumptions:

- All lines from the dataset terminate at a substation
- Each line connects precisely two substations
- Lines cannot "branch off" of other lines

Data cleaning begins with filtering out any substations that are not within reasonable proximity to a line in the dataset. These substations may be distribution substations that have been misidentified as transmission substations and do not play a part in the transmission network. Next, lines must be removed or separated in case they are combined in one object and are not considered if the voltage parameter is not present. In cases where lines are present but the terminating substation is missing from the dataset, it can be systematically added. Synthetic substations can also be added at points where lines "branch", splitting the line to which it may connect. These points may not be actual substations but may be one line tapping into another. These are still represented as buses in the PyPSA network and are converted to substations with a distinct identifying prefix.

3.3. Topology Construction

Network topology is the first step of creating the newly instantiated PyPSA network object. This process iteratively loops through the cleaned substations and lines datasets. Substations are assigned voltage levels based on the lines that are within a fixed proximity. For example, one substation may have the endpoint of two 230kV and one 69kV line, and would receive two voltage levels, 230kV and 69kV. Transformers are connected between the two buses with an assumed reactance and zero resistance. Transformers also are assumed to have infinite capacity. Lines are added next, with each connecting two substations. Electrical parameters for each line (resistance, reactance) are estimated statistically (Sadeghian et al., 2018) using distributions from (Athari & Wang, 2017), with different distributions for each voltage level. Any attributes from the geospatial lines database is carried through and added to the network object, aiding in geospatial analysis. A linear power flow is completed and any buses and lines which are completely islanded from the larger network are removed, and disregarded in the final network output. These could be features which are only connected with distribution infrastructure and don't have significant impact on the behavior of the network as a whole.

3.4. Assigning Power Plants

Power plants are modeled as generators in PyPSA and tied to the network at buses. The power plant information is iteratively added to the nearest bus within an adjustable geospatial tolerance. While some models perform optimal power flow to assign generator dispatch (Young et al., 2018), this requires significant overhead and limited value (Espejo et al., 2019). Instead, a statistical approach can be used through methods outlined in (Sadeghian et al., 2018). This works by assigning a dispatch factor i to each generator i. The dispatch factor is used to determine the generator's power output through the relationship in equation 1. The dispatch factor for all generators in a network is statistically correlated to the normalized generation capacity, which is a generator's capacity relative to the maximum generating capacity in the network (Sadeghian et al., 2018). The dispatch factor follows a statistical distribution shown in Figure 3, which is used for random assignment of generation dispatch factors in the model.

$$\alpha_i = P_i / P_i^{max} \tag{1}$$

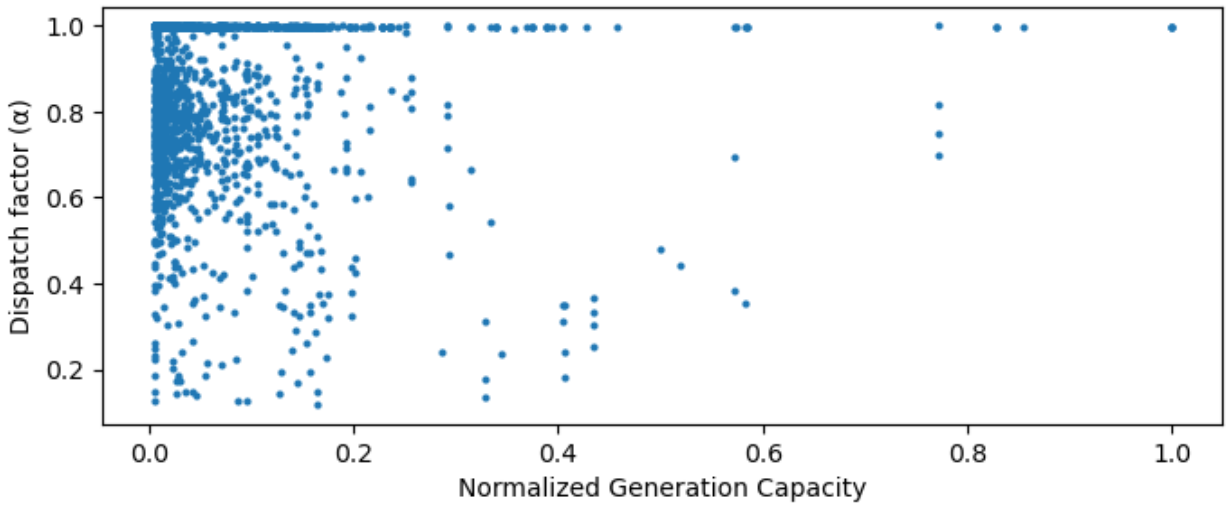

Figure 3: Generator dispatch factor vs normalized generation capacity from Birchfield WECC 10k bus network (Birchfield, Xu, et al., 2017) with methods from (Sadeghian et al., 2018)

3.5. Assigning Loads

The load assigned to each substation can be defined by looking at the electrical load of each substation service area. Service area polygons for each bus can traditionally be defined through Voronoi polygons (Held & Williamson, 2004; Sparks et al., 2023). In cases where buses of multiple voltage levels exist at a substation, the load is assigned to the bus with the lowest voltage. Spatially correlated load data are available in the US through geographical aggregated load data from NREL ComStock and ResStock (Landsman et al., 2024). This data provides aggregation by Public Microdata Use Area (PUMA). PUMAs are non-overlapping statistical geographic areas designated by the US Census Bureau, allowing access to highly detailed data while maintaining the privacy of the individual in the data (NREL, 2020). As a result, we must assume the load is homogeneously distributed within each PUMA. Peak load for each PUMA is then apportioned by area to the service area polygons drawn for each load substation. This results in a detailed modeling of the geospatial energy consumption across the US. Average hourly load profiles for each PUMA in the boundary of the sample region is shown in Figure 4.

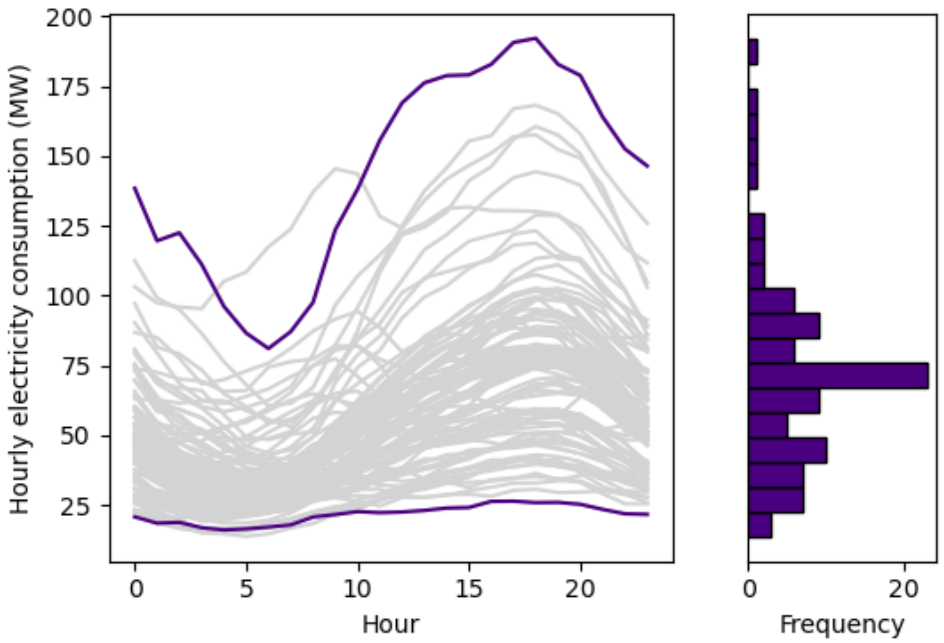

Figure 4: Daily load consumption by PUMA for the sample region (Landsman et al., 2024)

If detailed data on load distribution are not available, loads may be alternatively assigned through a linear correlation between population and load (Gegner et al., 2016). This allows any geospatially correlated population dataset to be used to assign population to each substation. After the population served by each transmission substation is found, load is assigned to each bus through the linear correlation. Figure 5 shows polygons representing the load service areas as well as generators.

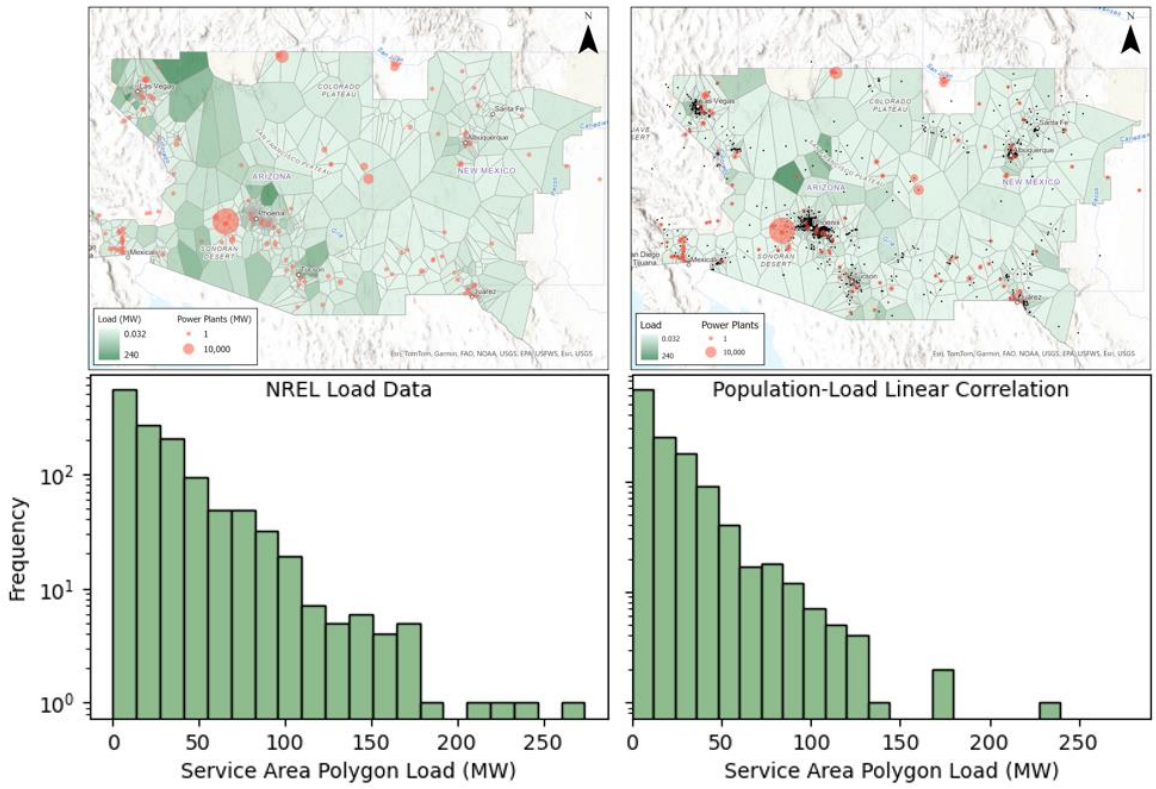

Figure 5: Placement of loads and power plants Histogram of service region load based on NREL ComStock and ResStock (left) and strict linear correlation with population (right)

3.6. Assigning Line Capacities

Line capacities can be synthetically assigned to the network through a process outlined by Sadeghian et al (Sadeghian et al., 2018). Similar to assigning generation dispatch factors, Sadeghian performs a statistical assignment of transmission gauge ratio for each line in the network. This uses the result from the linear power flow on the network after all loads and generators have been instantiated. Like the generation dispatch factor, the transmission line gauge ratio correlates with the normalized power flow through the line. This gauge factor relates to the line capacity through the relationship shown in equation 2. The relationship between the gauge ratio and the normalized power flow is shown in Figure 6.

$$\beta_l = F_l / F_l^{max} \tag{2}$$

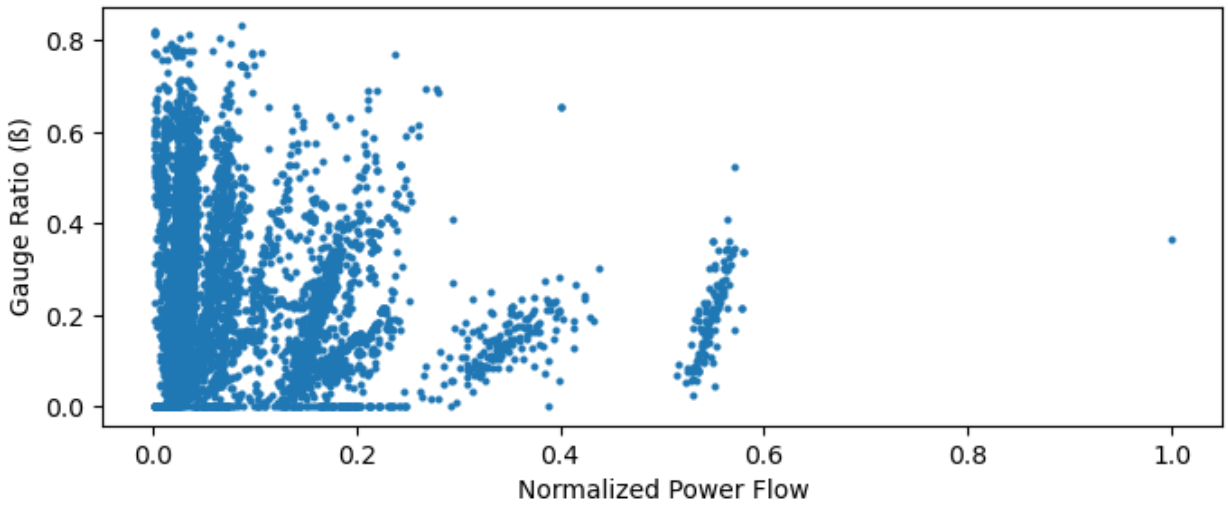

Figure 6: Transmission gauge ratio vs normalized power flow from Birchfield WECC 10k bus network (Birchfield, Xu, et al., 2017) with methods from (Sadeghian et al., 2018)

4. CASE STUDY: U.S. SOUTHWEST

The network estimation for the NERC AZ NM SNV subregion is shown in Figure 7. During the network construction process, lines are modeled directly between two substations only. This allows the network to be abstracted as seen in Figure 7 (right). Some regions that have lines which don't have a clear connection to the network are removed during the network construction process, so careful manual validation of the data should be used if these areas are of significant modeling importance. Descriptive statistics for the network are shown in Table 2 and are in line with the same parameters for fully synthetic networks (Birchfield, Xu, et al., 2017). Differences in generation values largely relate to the data used in the hybrid-synthetic model, which includes several recent utility-scale DER projects that may not have been captured when the fully synthetic model was developed.

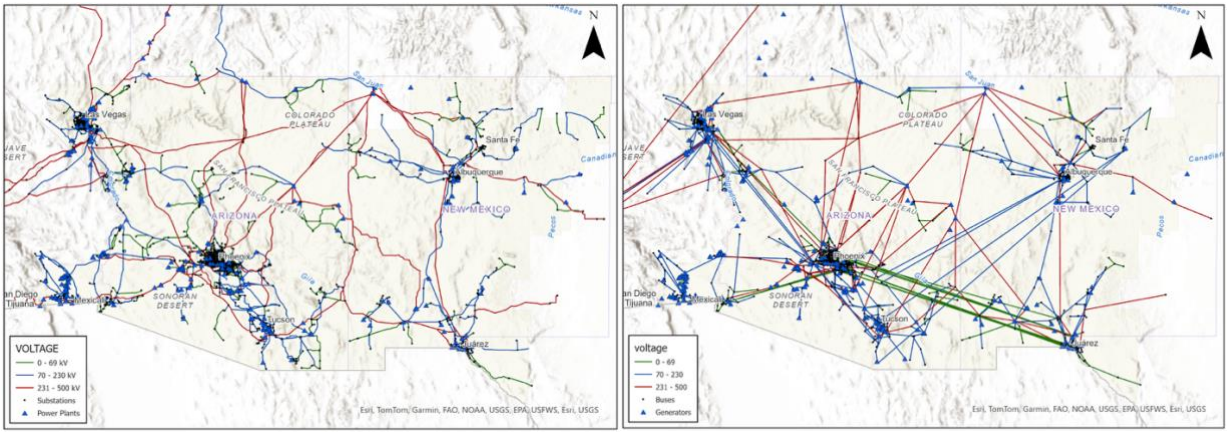

Figure 7: Geospatial data (left) and abstracted network output (right)

Table 2: Comparison of hybrid-synthetic and synthetic models

Statistic	Hybrid-Synthetic	Synthetic
Number of buses	1,380	1,747
Number of lines	1,471	1,940
Total load (MW)	24,676	28,824
Total generating capacity (MW)	57,539	36,916

5. CASE STUDY: CONTINENTAL U.S.

The process can also be generalized to the entire continental US, utilizing the entirety of the data available as shown in Figure 8. Being a linear power flow model, the model can also account for High Voltage DC (HVDC) lines that connect each interconnect. In reality, these lines have converter stations and additional hardware required for normal operation. In the linear power flow model, however, they are treated the same. Therefore, this approach can be used to gain a broad perspective for the functionality of the electric power system based on geospatially real data, incorporating the high-level dynamics of real power flow.

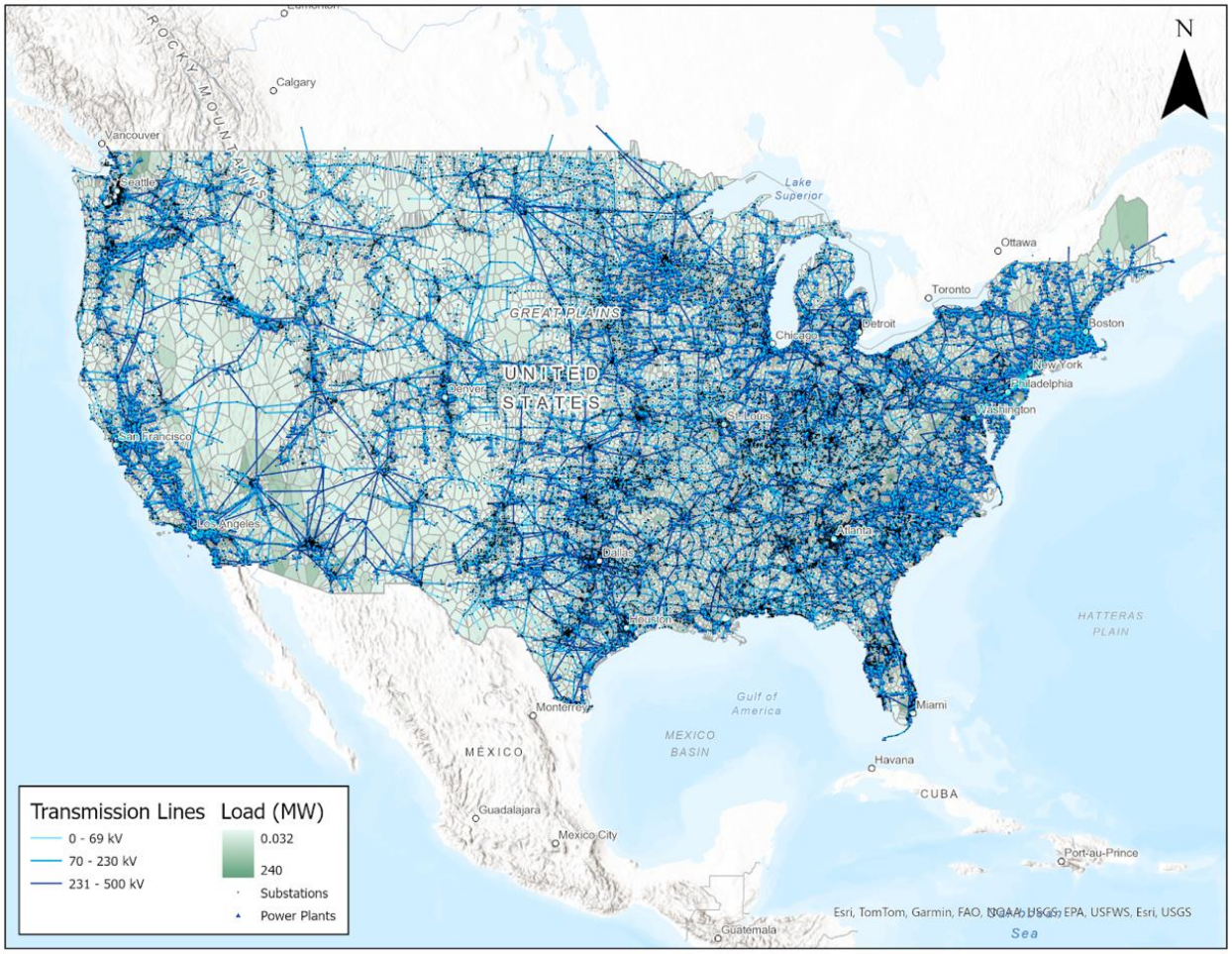

Figure 8: Network construction for the whole U.S.

6. REFERENCES

Abdel-Khalek, H., Schumm, L., Jalbout, E., Parzen, M., Schauß, C., & Fioriti, D. (2025). PyPSA-Earth sector-coupled: A global open-source multi-energy system model showcased for hydrogen applications in countries of the Global South. Applied Energy, 383, 125316. https://doi.org/10.1016/j.apenergy.2025.125316

Arianos, S., Bompard, E., Carbone, A., & Xue, F. (2025). Power grid vulnerability: A complex network approach.

Athari, M. H., & Wang, Z. (2017). Interdependence of Transmission Branch Parameters on the Voltage Levels. https://arxiv.org/abs/1709.06930

Birchfield, A. B., Gegner, K. M., Xu, T., Shetye, K. S., & Overbye, T. J. (2017). Statistical Considerations in the Creation of Realistic Synthetic Power Grids for Geomagnetic Disturbance Studies. IEEE Transactions on Power Systems, 32(2), 1502–1510. https://doi.org/10.1109/TPWRS.2016.2586460

Birchfield, A. B., Xu, T., Gegner, K. M., Shetye, K. S., & Overbye, T. J. (2017). Grid Structural Characteristics as Validation Criteria for Synthetic Networks. IEEE Transactions on Power Systems, 32(4), 3258–3265. https://doi.org/10.1109/TPWRS.2016.2616385

Birchfield, A. B., Xu, T., Shetye, K. S., & Overbye, T. J. (2018). Building synthetic power transmission networks of many voltage levels, spanning multiple areas. Proceedings of the Annual Hawaii International Conference on System Sciences, 2018-Janua, 2766–2774. https://doi.org/10.24251/hicss.2018.349

Borkowska, S., & Pokonieczny, K. (2022). Analysis of OpenStreetMap Data Quality for Selected Counties in Poland in Terms of Sustainable Development. Sustainability, 14(7), 3728. https://doi.org/10.3390/su14073728

Brown, T., Hörsch, J., & Schlachtberger, D. (2018). PyPSA: Python for power system analysis. Journal of Open Research Software, 6(1). https://doi.org/10.5334/jors.188

Espejo, R., Lumbreras, S., & Ramos, A. (2019). A Complex-Network Approach to the Generation of Synthetic Power Transmission Networks. IEEE Systems Journal, 13(3), 3050–3058. https://doi.org/10.1109/JSYST.2018.2865104

Farrag, M. A., Ali, K. M., & Omran, S. (2019). AC load flow based model for transmission expansion planning. Electric Power Systems Research, 171, 26–35. https://doi.org/10.1016/j.epsr.2019.02.006

Gegner, K. M., Birchfield, A. B., Xu, T., Shetye, K. S., & Overbye, T. J. (2016). A methodology for the creation of geographically realistic synthetic power flow models. 2016 IEEE Power and Energy Conference at Illinois, PECI 2016, 1–6. https://doi.org/10.1109/PECI.2016.7459256

Haklay, M. (2010). How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. Environment and Planning B: Planning and Design, 37(4), 682–703. https://doi.org/10.1068/b35097

Held, M., & Williamson, R. (2004). Creating electrical distribution boundaries using computational geometry. IEEE Power Engineering Society General Meeting, 2004., 2, 892–892. https://doi.org/10.1109/PES.2004.1372951

Helmrich, A., Chester, M., Miller, T. R., & Allenby, B. (2023). Lock-in: Origination and significance within infrastructure systems. Environ. Res.

Hörsch, J., Hofmann, F., Schlachtberger, D., & Brown, T. (2018). PyPSA-Eur: An open optimisation model of the European transmission system. Energy Strategy Reviews, 22, 207–215. https://doi.org/10.1016/j.esr.2018.08.012

Landsman, J., Bertolacini, M., Shah, A., Alberga, D., Andrews, D., Sontag, M., Sant, A. V., & Dahlhausen, M. (2024). Leveraging NREL's ResStock & ComStock Dataset to Evaluate Building Stock Electrification: Preprint. Renewable Energy.

Medjroubi, W., Müller, U. P., Scharf, M., Matke, C., & Kleinhans, D. (2017). Open Data in Power Grid Modelling: New Approaches Towards Transparent Grid Models. Energy Reports, 3, 14–21. https://doi.org/10.1016/j.egyr.2016.12.001

NREL. (2020). Final Criteria for Public Use Microdata Areas for the 2020 Census and the American Community Survey. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/pumas/2020pumas.html

Pandey, S., Patari, N., & Srivastava, A. K. (2019). Cognitive Flexibility of Power Grid Operator and Decision Making in Extreme Events. IEEE Power & Energy Society General Meeting (PESGM). https://doi.org/10.1109/PESGM40551.2019.8974015

Parzen, M., Abdel-Khalek, H., Fedotova, E., Mahmood, M., Frysztacki, M. M., Hampp, J., Franken, L., Schumm, L., Neumann, F., Poli, D., Kiprakis, A., & Fioriti, D. (2023). PyPSA-Earth. A new global open energy system optimization model demonstrated in Africa (Version 1.0.0) [Computer software]. https://doi.org/10.1016/j.apenergy.2023.121096

Sadeghian, H., Elyas, S. H., & Wang, Z. (2018, August 5). A Novel Algorithm for Statistical Assignment of Transmission Capacities in Synthetic Grid Modeling. 2018 IEEE PES General meeting.

Sadeghian, H., & Wang, Z. (2020). AutoSynGrid: A MATLAB-based toolkit for automatic generation of synthetic power grids. International Journal of Electrical Power and Energy Systems, 118. https://doi.org/10.1016/j.ijepes.2019.105757

Simpson-Porco, J. W. (2018). Lossy DC Power Flow. IEEE Transactions on Power Systems, 33(3), 2477–2485. https://doi.org/10.1109/TPWRS.2017.2749042

Sparks, R. M., Hoff, R., Johnson, N., & Chester, M. (2023). Identifying Cascading Failures on Synthetic Power Transmission Systems.

Stott, B., Jardim, J., & Alsac, O. (2009). DC Power Flow Revisited. IEEE Transactions on Power Systems, 24(3), 1290–1300. https://doi.org/10.1109/TPWRS.2009.2021235

Trpovski, A., & Hamacher, T. (2019). A Comparative Analysis of Transmission System Planning for Overhead and Underground Power Systems using AC and DC Power Flow. 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), 1–5. https://doi.org/10.1109/ISGTEurope.2019.8905510

U. S. Census Bureau. (2020). Total Population. https://data.census.gov/table/DECENNIALCD1182020.P1?q=Population+Total

U.S. Department of Homeland Security. (2025, March 10). Homeland Infrastructure Foundation Level Data—Transmission Lines. Transmission Lines. https://hifld-geoplatform.hub.arcgis.com/datasets/bd24d1a282c54428b024988d32578e59_0/explore

U.S. Energy Information Administration. (2020). NERC Regions [Dataset]. https://atlas.eia.gov/datasets/eia::nerc-regions/about

Villasana, R., Garver, L., & Salon, S. (1985). Transmission Network Planning Using Linear Programming. IEEE Transactions on Power Apparatus and Systems, PAS-104(2), 349–356. https://doi.org/10.1109/TPAS.1985.319049

Wiese, F., Schlecht, I., Bunke, W.-D., Gerbaulet, C., Hirth, L., Jahn, M., Kunz, F., Lorenz, C., Mühlenpfordt, J., Reimann, J., & Schill, W.-P. (2019). Open Power System Data – Frictionless data for electricity system modelling. Applied Energy, 236, 401–409. https://doi.org/10.1016/j.apenergy.2018.11.097

Young, S. J., Makarov, Y., Diao, R., Fan, R., Huang, R., OrBrien, J., Halappanavar, M., Vallem, M., & Huang, Z. H. (2018). Synthetic Power Grids from Real World Models. 2018 IEEE Power & Energy Society General Meeting (PESGM), 1–5. https://doi.org/10.1109/PESGM.2018.8585792