# A Geospatial, Cost-Aware Framework for Synthetic Distribution Grid Generation in Tribal Lands

Hanbyeol Shin
*Dept. of Civil, Environmental and Geodetic Engineering*
*The Ohio State University*
Columbus, OH, USA
shin.922@osu.edu

Abdollah Shafieezadeh
*Dept. of Civil, Environmental and Geodetic Engineering*
*The Ohio State University*
Columbus, OH, USA
shafieezadeh.1@osu.edu

***Abstract*—Reliable electricity access is fundamental for socioeconomic development, yet grid expansion in rural and remote regions presents significant planning challenges, often exacerbated by a lack of existing infrastructure data. This paper presents a novel framework for generating realistic synthetic power distribution grids in data-scarce regions using publicly available geospatial data. Our method constructs a cost-optimal network connecting buildings and critical facilities, then prunes it using a density-based algorithm. Crucially, this pruning is calibrated to match empirically observed electrification rates, ensuring the model reflects real-world service coverage. We validated the framework in the Navajo Nation, where the generated grid's topological properties (e.g., nodal degree, network density) closely matched a proprietary dataset of the as-built network. The resulting model serves as a practical and replicable tool for greenfield electrification planning, enabling better-informed infrastructure decisions in underserved communities.**

***Index Terms*—synthetic distribution network, underserved communities, grid topology, statistical validation, electrification**

## I. Introduction

Access to reliable electricity is a cornerstone of modern economic development, public health, and social equity. However, a significant portion of people remain underserved, particularly in rural, remote, and economically disadvantaged regions. Planning and deploying electrical distribution infrastructure in these areas is a complex and capital-intensive endeavor, requiring careful consideration of geography, population density, and economic constraints. To aid in this critical planning process, as well as for resilience studies and the analysis of grid modernization strategies, synthetic power grid models have become an indispensable tool. Standard synthetic grid generation tools, such as SMART-DS [1] and RNM-US [2], are primarily designed for fully developed regions and operate under the fundamental assumption of 100% building electrification. Consequently, these models lack the structural mechanisms to capture unserved rates in underserved communities. Furthermore, many existing models are either topologically abstract or lack adequate grounding in the real-world geographic and economic factors that shape grid design.

This material is based upon work supported by the U.S. Department of Energy's Solar Energy Technologies Office (SETO) under the award number DE-EE0010420 as well as the Lichtenstein Endowment at The Ohio State University.

This paper presents a novel, data-driven framework for generating realistic synthetic power distribution grids in data-scarce regions using publicly available geospatial data. Our methodology constructs a cost-optimal primary backbone connecting critical facilities, extends secondary service to all buildings, and employs a density-based pruning algorithm calibrated to match empirically observed electrification rates. We validate the framework through a case study in the Navajo Nation, where the generated grid's topological properties closely match the actual as-built network. The resulting model provides a practical, replicable tool for greenfield electrification planning, enabling better-informed infrastructure investment decisions in underserved communities.

## II. Methodology

### A. Assumptions and Data Requirements

Our model assumes the power grid is constrained to the road network graph $G_R = (V_R, E_R)$, reflecting real-world engineering practices where utilizing public road Right-of-Way (ROW) minimizes land acquisition costs, ensures maintenance access, and aligns infrastructure with demand centers. The model requires geospatial datasets including road networks, building footprints $B$, and locations of critical facilities (hospitals, schools, community centers). We assume that ensuring reliable electricity to these infrastructures is the first priority in grid design.

To translate this service-level target into a spatial criterion, we define a service-proximity threshold, $d_{\text{threshold}}$. This threshold is a deterministic value derived directly from the existing grid, $G_{\text{actual}}$, to represent the maximum distance at which a building is considered serviceable, corresponding to the target electrification rate $1 - U_{\text{target}}$.

$$d_{\text{threshold}} = \text{Percentile}(d(b, G_{\text{actual}}) \mid \forall b \in B, 1 - U_{\text{target}}) \quad (1)$$

where the function $d(b, G)$ calculates the Euclidean distance from building $b$ to its closest node on grid $G$.

In our framework, this constant is used to classify buildings relative to our generated final grid, $G_F$. A building $b$ is classified as:

$$\text{Status}(b) = \begin{cases} \text{Served} & \text{if } d(b, G_F) \le d_{\text{threshold}} \\ \text{Unserved} & \text{if } d(b, G_F) > d_{\text{threshold}} \end{cases} \quad (2)$$

979-8-3315-8138-1

### B. Primary Grid Generation

The primary grid, $G_P$, serves as the high-reliability backbone of the network. The generation process begins with the road network graph, $G_R = (V_R, E_R)$. As mentioned in section II-A, the first priority is connecting all essential facilities, which include the hospitals, schools, and community centers.

To integrate the facilities into the road network, each facility's geographic location is mapped to a vertex in the graph. Computationally, this process was implemented using the *Shapely* and *NetworkX* Python libraries to perform a nearest-neighbor search, identifying the vertex in $V_R$ with the minimum Euclidean distance to the facility's geographic coordinates. Applying this process to the facilities yields the set of critical vertices, $S \subset V_R$, which the primary grid must connect.

Each edge $e \in E_R$ is assigned a construction cost based on its physical length, $l(e)$, and its road type. The road type is quantified by a dimensionless cost coefficient, $c_{\text{class}}(e)$, which represents the relative expense and engineering difficulty of construction on different road types (e.g., assigning a lower cost to major highways than to unpaved local roads). In implementation, the edge weight used in routing is

$$w(e) = c_{class}(e) \cdot l(e) \tag{3}$$

with $l(e)$ measured in consistent metric units. The objective is to find a minimum-cost subgraph of $G_R$ that connects all vertices in $S$. This is achieved by first calculating the shortest weighted path cost, denoted as $C_p(u, v)$, for every pair of nodes $u, v \in S$. This cost represents the most economical route to build a power line between $u$ and $v$ and is found by running a shortest path algorithm on $G_R$ with weights $w(e)$. Formally:

$$C_p(u, v) = \min_{P_{uv} \subseteq G_R} \left( \sum_{e \in E(P_{uv})} w(e) \right) \tag{4}$$

where $P_{uv}$ is a path in $G_R$ connecting nodes $u$ and $v$, and $E(P_{uv})$ is the set of edges comprising that path.

With these pairwise costs computed, we determine the optimal backbone topology by modeling it as a Minimum Spanning Tree (MST) problem [3]. Specifically, we use Kruskal's algorithm [4], a greedy algorithm that is proven to find the MST by iteratively adding the lowest-cost edge that does not form a cycle. We consider an abstract, complete graph where the vertices are the set $S$ and the weight of an edge between any two vertices $u, v \in S$ is the pre-computed path cost $C_p(u, v)$. Let $Y$ be the set of edges in the final spanning tree. Each edge is an unordered pair of vertices, denoted $\{u, v\}$, representing the decision to connect the critical nodes $u$ and $v$. The optimization problem is to find the set of edges $Y$ that forms a tree on the vertices $S$ while minimizing the total cost:

$$\begin{aligned} \underset{Y}{\text{minimize}} \quad & \sum_{\{u,v\} \in Y} C_p(u, v) \\ \text{subject to} \quad & \text{The graph } (S, Y) \text{ is a tree.} \end{aligned} \tag{5}$$

This choice yields a cost-optimal, acyclic, and connected backbone over $S$.

The final primary grid, $G_P = (V_P, E_P)$, is constructed by taking the union of all road segments comprising the shortest paths in $Y$, ensuring a connected, cost-optimal backbone linking all critical facilities.

### C. Secondary Grid Generation

The secondary grid extends the primary backbone to provide connectivity to all buildings. This process connects each building in the set $B$ to $G_P$. Operationally, secondary lines are routed along the road graph using the same edge weights defined in equation (3), to preserve economic consistency with the backbone construction.

For each building $b \in B$, represented as a geographic coordinate, we first identify its corresponding vertex on the road network, $v_b \in V_R$ by snapping its coordinates to the nearest node in $G_R$.

Next, we compute the optimal connection path, $P_b$, for each building. This path represents the most cost-effective route from the building's node $v_b$ to $G_P$. For each building, $P_b$ is the solution to the following shortest path optimization problem:

$$\begin{aligned} \underset{P \subseteq G_R}{\text{minimize}} \quad & \sum_{e \in E(P)} w(e) \\ \text{subject to} \quad & P \text{ is a simple path,} \\ & \partial P \cap v_b \neq \emptyset, \\ & \partial P \cap V_P \neq \emptyset. \end{aligned} \tag{6}$$

The objective function minimizes the construction cost over the set of edges in path $P$, denoted $E(P)$. The constraints formally define the path's connectivity. The path $P$ should be a simple path, which does not visit any vertex more than once, thereby preventing non-optimal loops. Here, $\partial P$ denotes the boundary of the path $P$, which is a set of vertices containing its start and end points. The constraints thus mandate that one endpoint of the path must be the building's node $v_b$, and the other endpoint must be a vertex within the primary grid's vertex set, $V_P$. This problem is solved for every building $b \in B$ to find its respective connection path $P_b$. The complete secondary grid, $G_S = (V_S, E_S)$, is constructed by forming the union of $G_P$ with all the newly computed paths. The resulting grid, $G_S$, is a fully connected network that serves as the maximal candidate for the subsequent calibration stage.

### D. Grid Pruning & Calibration

The secondary grid, $G_S = (V_S, E_S)$, represents a network with maximum possible connectivity, connecting every building to the primary grid backbone. To reflect the economic and practical constraints of grid deployment, we introduce an algorithm that prunes segments that are not economically viable. This process is governed by a key hyperparameter: the minimum density threshold, $k$, defined in units of buildings per kilometer. The algorithm iteratively identifies and removes terminal grid segments (dead-end branches) of a fixed 1 km length that serve a number of buildings below this threshold $k$.

This pruning is restricted to leaf paths, ensuring that backbone connectivity is preserved.

Since the optimal value of $k$ is not known a priori, we calibrate it by comparing the model's output to the ground-truth unserved rate, $U_{\text{target}}$. We define a set of candidate density values, $K$. For each candidate $k \in K$, the pruning algorithm is executed on the initial secondary grid $G_S$ to produce a candidate final grid, denoted $G_F(k)$. In our experiments, $K$ is taken as a small integer grid of plausible densities, enabling efficient calibration by enumeration.

For each resulting grid $G_F(k)$, we calculate its corresponding unserved rate, $U(k)$. This is achieved by applying the predefined service distance threshold, $d_{\text{threshold}}$ from equation (1), and the classification criterion, equation (2). The unserved rate is the fraction of buildings in the total set $B$ whose minimum Euclidean distance to the grid $G_F(k)$ exceeds this threshold:

$$U(k) = \frac{|\{b \in B \mid d(b, G_F(k)) > d_{\text{threshold}}\}|}{|B|} \tag{7}$$

This metric directly links pruning-induced sparsity to service accessibility at the building level.

The optimal density threshold, $k^*$, is then selected as the candidate value from $K$ that minimizes the absolute difference between the model's calculated unserved rate and the target rate:

$$k^* = \arg\min_{k \in K} |U(k) - U_{\text{target}}| \tag{8}$$

The final, calibrated synthetic grid, $G_F$, is the grid generated using this optimal threshold:

$$G_F = G_F(k^*) \tag{9}$$

This calibration step ensures that the final topology embodies the empirically observed electrification level while remaining economically feasible. The complete procedure for this step, from pruning to calibration, is formally detailed in Algorithm 1.

**Algorithm 1** Algorithm for Grid Pruning and Calibration

**Input:** Initial grid $G_S$, Buildings $B$, Target unserved rate $U_{\text{target}}$
**Parameters:** Set of density thresholds $K$, Service distance $d_{\text{threshold}}$
**Output:** Final calibrated grid $G_F$

**Function** CalculateUnservedRate($G_{\text{cand}}, B, d_{\text{threshold}}$):
    $B_{\text{unserved}} \leftarrow \{b \in B \mid d(b, G_{\text{cand}}) > d_{\text{threshold}}\}$
    **return** $|B_{\text{unserved}}|/|B|$

**for** each density threshold $k \in K$ **do**
    $G_{\text{candidate}} \leftarrow G_S$ {Start with a copy of the initial grid}
    **while** true **do**
        {Iterate until the grid stabilizes}
        Find a leaf branch $P$ in $G_{\text{candidate}}$ with building density less than $k$.
        **if** such a branch $P$ was found **then**
            Remove branch $P$ from $G_{\text{candidate}}$.
        **else**
            **break** {No more branches to prune; grid is stable}
        **end if**
    **end while**
    Store the stabilized grid as $G_F(k)$.
    $U(k) \leftarrow$ CalculateUnservedRate($G_F(k), B, d_{\text{threshold}}$)
**end for**
Find the best threshold: $k^* \leftarrow \arg\min_{k \in K} |U(k) - U_{\text{target}}|$.
**return** The grid corresponding to the best threshold, $G_F(k^*)$.

## III. Validation

### A. Case Study

To validate our grid generation framework, we apply it to a specific region within the Navajo Nation. This case study area spans approximately 2,600 square kilometers and is home to an estimated population of 10,000 residents. The selection of this region is primarily motivated by the availability of a dataset to the research team, detailing the actual, as-built electrical distribution grid. Because actual distribution grid topologies are classified as critical energy infrastructure information and are generally proprietary, this region represents the sole territory where ground-truth network data was made available to the research team, providing a unique and invaluable benchmark for validating the topological and spatial accuracy of our generated network.

A key parameter for our model is the target unserved building rate, $U_{\text{target}}$. For this region, we adopt the value of 0.21, which reflects the finding that approximately 21% of homes on the Navajo Reservation lack access to electricity [5]. This target rate is also used to define the proximity threshold for classifying a building as "served." Following equation (1), we calculated the building-to-grid connection distance, $d_{\text{threshold}}$, by finding the percentile of distances from all buildings to the actual grid, $G_{\text{actual}}$, that corresponds to the target served rate. This yielded a $d_{\text{threshold}}$ of 224 m, which serves as the benchmark for evaluating our generated grid.

The geospatial inputs required by our model were compiled from publicly available sources. The road network shapefile was derived from the U.S. Census Bureau [6]. Building footprints were sourced from the FEMA USA Structures [7]. Essential facilities, including hospitals, schools, and chapter houses, were also collected from publicly available sources [8]–[10]. Finally, the location of the distribution substation within the area was identified using high-resolution satellite imagery and designated as the root node for our grid generation algorithm.

To ensure the generated grid, $G_F$, realistically reflects the service levels of the actual territory, we calibrate the pruning parameter, $k$. This parameter, a building density threshold (buildings/km), balances the cost of grid extension against the benefit of connecting more buildings. To assess the robustness of this parameter, we performed a sensitivity analysis by varying $k$ and observing the impact on the grid's unserved rate, $U(k)$. Following Algorithm 1, our analysis showed that a threshold of $k = 3$ resulted in 12.45% unserved rate, while $k = 5$ yielded 23.63%. Ultimately, $k = 4$ yielded an 18.52% unserved rate, the closest match to the 21% target. Accordingly, we selected the optimal threshold of $k^* = 4$ buildings/km for our case study.

979-8-3315-8138-1

TABLE I
COMPARISON OF STATISTICAL METRICS FOR THE GENERATED AND ACTUAL GRIDS

| Metric | Generated Grid | Actual Grid |
|---|---|---|
| Total Length (km) | 741.86 | 858.84 |
| Average Degree | 2.0071 | 2.0886 |
| Average Betweenness | 0.0819 | 0.0220 |
| Network Density | 0.0002 | 0.0002 |

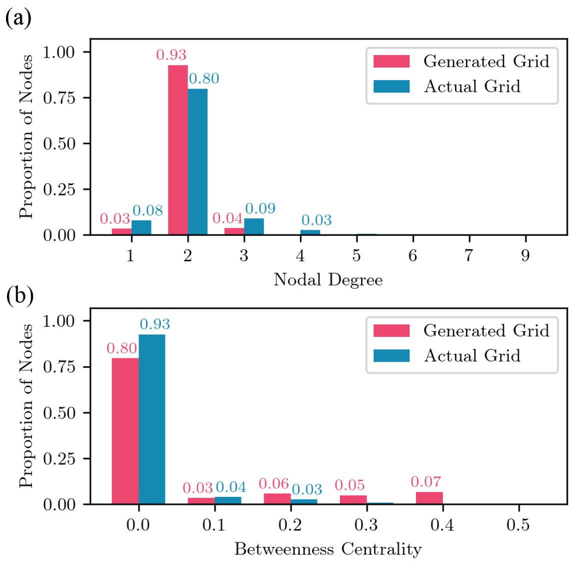


Fig. 1. Comparison of Topological Properties: (a) Nodal Degree Distributions; (b) Betweenness Centrality Distributions

### B. Statistical Validation

To evaluate the topological fidelity of our generated grid, $G_F$, we compare it against the actual grid using several key graph-theoretic metrics. Table I summarizes the primary statistics for both networks. Despite the generated grid being smaller in scale, with approximately 13.6% shorter total line length, the model successfully replicates key structural properties. The average nodal degree (i.e., the number of edges connected to a node) and network density (i.e., $\frac{2|E|}{|V|(|V|-1)}$, where $|E|$ is the number of edges and $|V|$ is the number of nodes) of the generated grid show strong agreement with those of the actual grid. This indicates that our model accurately captures the fundamental local connectivity and global sparsity of the real-world network, which are critical for realistic power flow and reliability analyses.

To delve deeper into the network structure, we compare the distributions of two central topological properties: nodal degree and betweenness centrality. The nodal degree, the most basic measure of local connectivity, is shown in Fig. 1(a). The degree distributions of both grids are remarkably similar. Both networks are overwhelmingly characterized by nodes of degree 2, a typical feature of rural electrical grids that contain long, chain-like feeder lines. While minor variations exist, with the generated grid showing a slightly higher concentration of degree-2 nodes, the overall profile is well-preserved. This demonstrates the model's ability to replicate the fundamental local topology of the target network.

A more global measure of a node's importance is its betweenness centrality, which quantifies how often a node acts as a bridge along the shortest paths between other nodes. Formally, the betweenness centrality $c_B(v)$ of a node $v$ is given by [11]:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{10}$$

where $V$ is the set of all nodes, $\sigma(s,t)$ is the number of shortest paths between nodes $s$ and $t$, and $\sigma(s,t|v)$ is the number of those shortest paths that pass through $v$. Fig. 1(b) compares the normalized betweenness centrality distributions. The comparison of betweenness centrality distributions, shown in Fig. 1(b), further underscores the model's success. Both distributions are heavily skewed, with the vast majority of nodes in both networks (0.80 in the generated grid, 0.93 in the actual grid) having centrality scores near zero. This is a hallmark of sparse, dendritic networks and indicates that the generated grid correctly reproduces this essential global property.

The generated grid does exhibit a slightly heavier tail, with a larger proportion of nodes having low-to-moderate centrality. This leads to a higher average betweenness value (Table I). However, the goal is to generate a realistic, not identical, network. The strong qualitative agreement in the distributional shapes seen in Fig. 1(b) is far more significant than the difference in a single summary statistic. It confirms that our model produces a network with a realistic centrality profile, capturing the general character of how importance is distributed across the nodes.

In summary, the validation confirms that the generative model produces networks with high topological fidelity. By successfully replicating the statistical distribution of both local (nodal degree) and global (betweenness centrality) metrics, the model demonstrates its capability to generate synthetic grids that are structurally realistic and suitable for subsequent network analysis.

## IV. APPLICATION

To demonstrate the practical utility of our framework, we apply the calibrated model to a new region where the actual grid topology is unknown. The selected application area is a region within the Navajo Nation spanning seven chapters and containing approximately 4,600 buildings. The absence of verified grid data for this territory makes it an ideal test case for our predictive generation methodology.

We assume that the service distance threshold $d_{\text{threshold}}$ of 224 m and optimal density threshold $k^*$ of 4 buildings/km, calibrated from one region of the Navajo Nation, is transferable to other areas within the reservation given similar terrain, utility practices, and housing density patterns. Using publicly available geospatial data, we generated a predictive distribution grid for the territory. The resulting network is illustrated

979-8-3315-8138-1

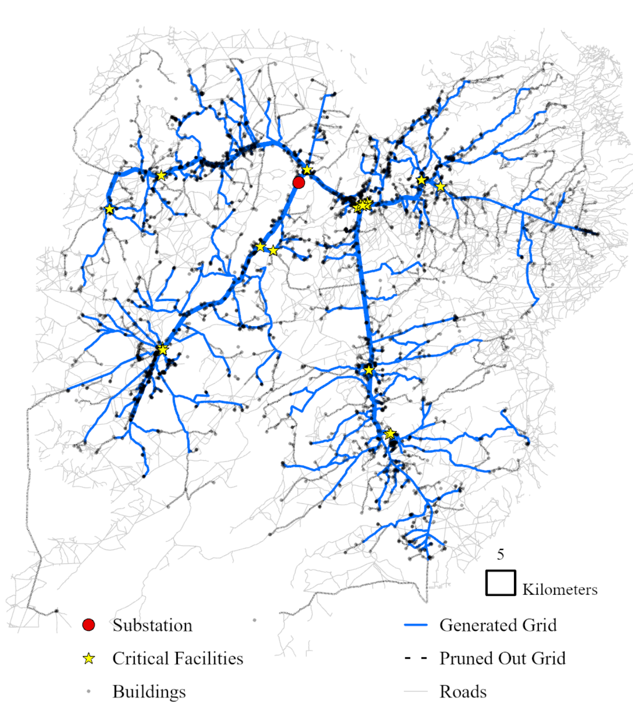


Fig. 2. Generated distribution grid for a region in the Navajo Nation. The final grid (blue) is shown alongside the pruned-out segments (black dashed), buildings, roads, critical facilities, and the primary substation.

in Fig. 2. The final grid is shown in blue, connecting the substation (red circle), critical facilities (yellow stars) and thousands of buildings (black dots), while the black dashed lines represent potential extensions that were pruned by the algorithm for not meeting the cost-benefit threshold. It is important to emphasize that this parameter transferability is strictly valid for regions sharing similar socio-economic and geographic constraints. Application of this framework to fundamentally different topographies, such as dense urban environments, would inherently require re-calibration of the model parameters.

Classifying buildings using equation (2), we estimate the electrification rate for the region at 75%. This figure aligns well with the reported 79% rate for the broader Navajo Nation [5]. This application demonstrates that the framework can generate plausible, geographically explicit grids and provide realistic service coverage estimates for infrastructure planning and policy in data-scarce environments.

## V. Conclusion

This paper addressed the critical challenge of designing cost-effective electrical grid extensions for underserved communities. We introduced a systematic, three-stage optimization framework that leverages real-world geospatial data to generate a practical grid topology. The model first establishes a primary backbone connecting critical community facilities via a MST and subsequently connects individual households through optimized paths along the existing road network. The primary contribution of this work is a scalable and data-driven planning tool that provides an objective, cost-optimized baseline for electrification projects, replacing subjective or manual design processes with a computationally efficient and replicable methodology. Crucially, by explicitly simulating where grid extensions become economically unviable, the model allows planners to pinpoint electrification gaps, identifying who remains unserved and where alternative off-grid solutions are needed.

The proposed framework serves as a vital decision-support tool for policymakers, utilities, and Tribal governments, enabling them to strategically allocate resources and accelerate the deployment of essential infrastructure. While our current model provides a robust foundation based on network distance, future work should focus on integrating more sophisticated cost functions that account for terrain, land ownership, and construction complexities. Furthermore, a significant extension would be to incorporate the optimal placement of distributed energy resources, such as solar-plus-storage systems, to evaluate hybrid grid-extension and microgrid solutions. Ultimately, advancing such quantitative planning tools is essential for engineering a future where reliable and affordable energy access is equitable for all communities.

## References


[1] V. Krishnan, B. Bugbee, T. Elgindy, C. Mateo, P. Duenas, F. Postigo, J.-S. Lacroix, T. G. San Roman, and B. Palmintier, "Validation of synthetic us electric power distribution system data sets," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4477–4489, 2020.

[2] C. Mateo, F. Postigo, F. de Cuadra, T. G. San Roman, T. Elgindy, P. Duenas, B.-M. Hodge, V. Krishnan, and B. Palmintier, "Building large-scale us synthetic electric distribution system models," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5301–5313, 2020.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2022.

[4] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical society*, vol. 7, no. 1, pp. 48–50, 1956.

[5] U.S. Department of Energy, Office of Indian Energy Policy and Programs, "Tribal electricity access and reliability report to congress august 2023," U.S. Department of Energy, Report to Congress, August 2023. [Online]. Available: https://www.energy.gov/indianenergy/articles/tribal-electricity-access-and-reliability-report-congress-august-2023

[6] U.S. Census Bureau, "TIGER/Line Shapefiles," https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html, 2024, accessed: May 23, 2024.

[7] Federal Emergency Management Agency (FEMA), "USA Structures," https://gis-fema.hub.arcgis.com/pages/usa-structures, 2024, accessed: May 23, 2024.

[8] Indian Health Service (IHS), "Indian Health Service Facility Locations," https://www.ihs.gov/locations/, 2023, data updated June 15, 2023. Accessed: April 28, 2024.

[9] New Mexico Public Education Department, "New Mexico Public School Locations," 2022, the dataset was accessed by the authors on Feb 11, 2024. The original hosting webpage is no longer available as of Nov 2025.

[10] Navajo Nation Division of Community Development, "Navajo Nation ASC Map," https://www.arcgis.com/home/item.html?id=db69242bd4db4d52bfae9b9d713b775e, 2024, hosted on ArcGIS. Accessed: May 23, 2024.

[11] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation," *Social networks*, vol. 30, no. 2, pp. 136–145, 2008.


979-8-3315-8138-1