*14th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP14*
*Dublin, Ireland, July 9-13, 2023*

# Power Distribution Network Type Classification using a Machine Learning Approach

Ming Hong
*Data Scientist, One Concern Inc., Menlo Park, USA*

Chengwei Zhai
*Staff Data Scientist, One Concern Inc., Menlo Park, USA*

Youngsuk Kim
*Sr. Data Science Manager, One Concern Inc., Menlo Park, USA*

Shabaz Patel
*VP of Data Science, One Concern Inc., Menlo Park, USA*

ABSTRACT: Power distribution network vulnerability has been a critical component in measuring community resilience under natural disasters. Given overhead power lines exposed to extreme weather events are susceptible to large-scale damage and failure, it is imperative to identify if the power distribution network types are overhead or underground as part of the power outage prediction. As such data are not publicly available, we propose the application of machine learning techniques for power distribution network type classification. The purpose of this article is to improve the accuracy and generalizability of the power network type classification model proposed originally by Zhai et al. (2020). Given that most power distribution networks follow road networks, we labeled the distribution network type for over 60,000 selected road locations across major cities in the United States. We then combine the power distribution network type dataset with nearby building characteristics, road types, probabilistic hazard maps, and geographical location information to form a complete dataset for the training of the network type classifier. We predict the network type at the building level, then aggregate the predictions back to individual road segments. We demonstrate the performance using different machine learning models, feature combinations, and aggregation methods. As a result, the best performance model is able to predict the existence of an overhead system with a testing accuracy of over 75% and F1 score over 0.74. We conclude that our machine learning model is an effective and efficient tool for power distribution network type classification, which can be further applied to evaluate distribution network damage under natural disasters.

## 1. INTRODUCTION

All around the world, extreme hazard events such as flooding, tornadoes, and hurricanes have caused widespread power network failure for distribution systems. When Hurricane Irma struck Florida on September 10, 2017, 6.7 million customers experienced power outages, which represents almost one-third of the state population, and some areas had outages longer than a week (Chakalian et al., 2019). Moreover, large-scale damage to the power distribution network can have even more severe impacts on the economy. In 2012, Hurricane Sandy caused power outages for over 8.5 million people, which contributed to over $7 billion in economic losses (Nateghi et al., 2016). Overhead systems are known to be vulnerable to wind hazards (Waseem and Manshadi, 2020), while underground systems

are potentially susceptible to flood erosion (Miura et al., 2021). To understand the potential power distribution infrastructure vulnerability to different types of disaster events, it is pivotal to identify if the power system is overhead or underground in power system risk analysis and reliability assessment.

Currently, the majority of research on overhead distribution power lines centers on identifying and addressing vulnerabilities due to different types of disaster events. For instance, Zhou et al. (2006) proposed two methods to predict weather-related power system failure related to overhead power lines using Poisson regression and Bayesian network. To evaluate the likelihood of distribution system interruption at a large scale for certain types of disasters, an important first step after obtaining the power distribution network topology is to identify the power line type. However, there are very limited studies focusing on the classification of overhead and underground power networks. Such information is necessary not only because the overhead and underground power systems are prone to different types of damage. It is also essential to come to accurate power outage estimations at the hyper-local level. Power distribution networks are usually classified information, and the data is not publicly available. This is pointed out previously by Zhai et al. (2020) as one of the major drawbacks blocking simulation-based power outage analysis. In this study we have used the road segments as a proxy of the power distribution network because we do not have access to sensitive data.

Zhai et al. (2020) previously proposed the methodology of generating synthetic power networks and using a random forest machine learning model to infer overhead and underground power networks given nearby building features such as year built, value, and total square area for Columbus, OH. The paper showed very promising results for a given city, yet it might not be applicable for a larger spatial scale prediction. There are also several existing studies focusing on overhead power distribution line recognition from images or point cloud data. Roussel et al. (2021) published an algorithm used to recognize transmission towers and overhead power lines from 3D point clouds. Prates et al. (2019) used Convolutional Neural Networks (CNN) to detect the insulators in overhead power lines. Both approaches require using laser scans or photo-taking at power line locations, which can be more accurate but expensive to scale to a nationwide and global level.

In this study, we further improve upon the method mentioned by Zhai et al. (2021) on using machine learning models to predict the existence of overhead power networks. We aim to use a nationwide dataset and include more features than year-built value and total square area to improve the classification model's accuracy and generalizability for the continental US.

## 2. METHODOLOGY

The characteristic and topology of a power distribution network are often not publicly available in the U.S.. Given the necessity for both power lines and roads to reach individual buildings, this study assumes that the road network can serve as a proxy of the power distribution network, and the power distribution network type can be inferred from the surrounding building characteristics. The machine learning model trained from this study is specifically for the U.S., and the same methodology can be generalized to other countries.

### *2.1. Datasets*

The backbone of this study is the ground truth power network type dataset that was gathered using map services. By manually looking at over 60,000 road segments across most major cities in the U.S. to in order to ensure generalization across the entire country, the distribution network types are recorded, which is the machine learning model's response variable. The power network type can have three different labels: "True" means overhead power lines are observed in the street view; "False" means no overhead power lines are observed; "Unknown" means StreetView is not available. In the following steps, the locations with "Unknown" labels are excluded. Using the described ground truth dataset, the target for the trained machine learning model is to predict if an overhead power network exists at a given location. Figure 1 visualizes the

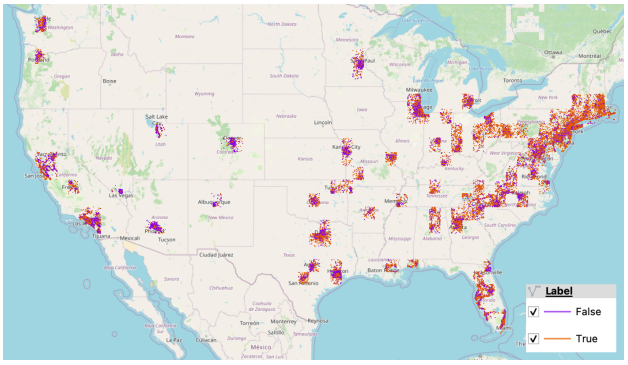


*Figure 1: Ground truth road segment points on map*

locations of the ground truth power network type dataset.

There exist various indicators that can be used to determine whether a given region or community is dominated by an overhead power system. Building-related features are the first to be considered. The year of construction of a building can be a strong indicator. In general, newly constructed communities are more likely to have underground power lines. Additionally, the cost of installation is a significant factor in determining power line types in a particular region. The installation of underground power lines incurs higher costs than overhead power lines. Thus, they are more prevalent in upscale neighborhoods, commercial districts or city centers. Other than value and year-built information, other building characteristics of a neighborhood can potentially be dependent variables to power line types, such as the number of floors, construction type, occupancy type, etc. Furthermore, the land use type of a region can also be an important factor. For example, a city's downtown area often prioritizes the installation of underground power lines. To generalize the assumption, the building density of an area or the classification of urban or rural regions can be a contributing factor to the power network type.

On the other hand, a number of non-building related factors also serve as potential features for power line type classification, such as exposure to hazards, road type, and geographical location. From the perspective of utility companies, maintenance of underground power lines presents additional challenges because excavation is needed for replacements or repairs when damages occur. Consequently, in regions exposed to repeated flooding events, it might be more difficult to keep the power lines underground. On the contrary, a region with frequent wind exposure might be more inclined to install underground power lines, as overhead power lines are vulnerable to damage caused by falling trees or debris. Additionally, the type of road itself might determine the power line type. It is likely that underground power lines are installed along the primary road, such as highways or expressways, while overhead power lines are installed along the local roads. Lastly, the local government legislature or geographical location could influence the power line types. The level of support for installing underground power lines may vary among local governments. Furthermore, mountainous regions mostly have overhead power lines because of low population density and easy repair.

Four additional datasets are selected to provide independent variables to the machine learning model. As previously discussed in the introduction section, surrounding building characteristics, road types, hazard exposure, and classification of urban or rural settings all could contribute to the inference of power line types. To capture the potential influence of local government legislation and geographic location, the political stance of the state based on the past 13 presidential election votes and the time zone the buildings are located in are also introduced as independent variables. The list below summarizes the important datasets used in this study.

1. Power line type dataset: Ground truth dataset with power line types manually labeled at over 60,000 road segments

2. Building characteristics dataset: Contains building characteristics variables including but not limited to total area square meters, year built, value, number of floors, occupancy, and vs30 value at the building locations

3. Road type dataset: Contains MTFCC_Code variable that categorizes road types (U.S. Geological Survey, National Geospatial Technical Operations Center, 2022)
   The most common road type in the roads

dataset is S1400: local neighborhood road, rural road, city street

4. Wind and flood hazard maps: Contains wind speed, storm surge depth, and inland flooding depth at different return periods. Hazard intensity values at a return period of 500 years are used for interpolation at individual building locations

5. Urban / Rural classification: Contains geometries corresponding to urban areas, urban centers, and rural regions (U.S. Census Bureau, 2017)

### *2.2. Power distribution network classification framework*

We begin compiling the complete dataset by converting all datasets to the individual building level by applying the 250m buffer. Then split the complete dataset into training and testing datasets. We then train the machine learning models, aggregate the building-level prediction result to road segment level, and evaluate the model performances at aggregated road level.

#### *2.2.1. Data Preprocessing*

The first step of the process is to build a complete dataset using all datasets discussed above. For each road segment, a 1,000-meter buffer is applied to filter for buildings within the buffer, resulting in over 18 million buildings included in the complete dataset. The distance between the central road segment and each building is also recorded as dist_to_road. The dist_to_road feature becomes a useful feature in the later steps to further filter for buildings. The power network type labels and road types are applied to all buildings within the buffer based on the central road segment label; building level features such as the number of floors or value of the property are left as is; other features such as urban/rural classification are spatially joined based on individual building locations. GeoPandas in python as well as QGIS, was used in this step for geographical processing.

#### *2.2.2. Feature Selection*

The complete dataset has 14 variables in total, including both numerical and categorical ones shown in the list below.

1. year_built: year the building is built
2. value: value of the building
3. total_area_sq_meters: total square meters of the building
4. floors: number of floors of the building
5. occ: occupancy type of the building:
   SFD: Single Family Home, MFD: Multi Family Home, IND: Industrial Building, COM: Commercial Building
6. num_bldg_nearby: Number of buildings exist around each road within 1000-meter buffer
7. MTFCC_code: MTFCC road type code
8. VS30: VS30 at the building location (the time-averaged shear-wave velocity to 30 m depth) to account for seismic hazard
9. wind_speed_RP500: Wind speed at return period 500 years at the building location to account for wind hazard
10. SS_RP500: Storm surge depth at return period 500 years at the building location
11. FL_RP500: Inland flooding depth at return period 500 years at the building location
12. UATYP10: Urban / Rural classification with 3 possible values; U: Urban Area, C: Urban Center, R: Rural
13. time_zone: Time zone the building is located in: Pacific, Mountain, Central, or Eastern
14. political_stance: Assigned to buildings based on the State it's located in. Categorized into 2 groups: red or blue, depending on the past 13 presidential election results

The unit count is another building characteristic feature. However, it is excluded after we conducted basic exploratory data analysis. To identify multicollinearity in the variables, a confusion matrix is used on the numerical variables and shown in figure 2 below. From the confusion matrix, we have identified that the two most correlated features are unit counts of the building, and the building's total

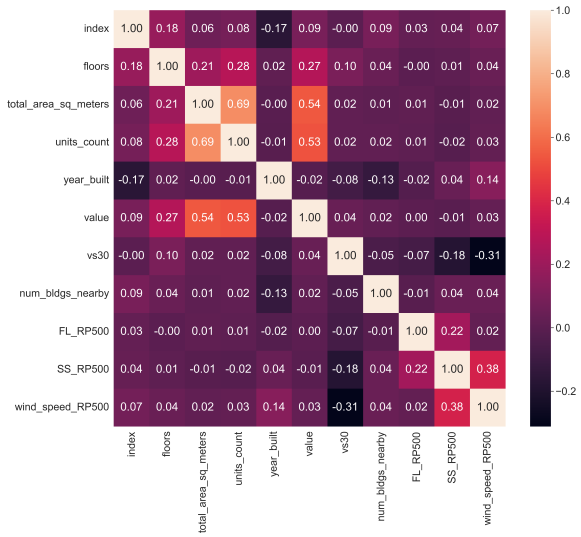


*Figure 2: Covariance matrix of continuous variables*

square footage, which makes sense. We have used a covariance of 0.6 as a threshold to exclude features with highly correlated variables. The unit count is hence removed from the selected features.

### *2.2.3. Data Versions*

In the complete dataset, we have selected 14 features in total. However, the scalability of the trained machine learning modeling also plays a part in the feature selection. The U.S. has over 100 million buildings in total, and preparing the complete dataset would require major computational power. For example, the building characteristics related variables are already in the building dataset and do not take extra effort to process. The hazard maps related variables, as well as urban/rural classification variables, are more difficult to obtain for a new set of buildings because they would require interpolation at each building location for the hazard values, and the urban/rural classification variables require spatial geoprocessing that takes a long time because of the complicated shapefile geometry. Considering different variables have different levels of difficulty to obtain at the building level, we have proposed several data versions with different sets of features ranging from easy to more difficult levels of preprocessing. The table 1 below shows the summary of features included in each data version.

*Table 1: Summary of data versions.*

| Data Version | Included Features |
|---|---|
| V1 | year_built, total_area_sq_meters |
| V2 | year_built, total_area_sq_meters, value, floors, occ |
| V3 | year_built, total_area_sq_meters, value, floors, occ, num_bldgs_nearby, VS30, FL_RP500, SS_RP500, wind_speed_RP500, MTFCC_CODE, UATYP10 |
| V4 | year_built, total_area_sq_meters, value, floors, occ, num_bldgs_nearby, VS30, FL_RP500, SS_RP500, wind_speed_RP500, MTFCC_CODE, UATYP10, time_zone, political_stance |

### *2.2.4. Buildings Filtering*

The second step is to determine the buffer distance that needs to be applied to further filter the buildings surrounding the road segments. To decide on the optimal buffer distance, an experiment was set up to compare classification performances with buffer distances 50m, 100m, 250m, and 500m. The same models were used to train on a filtered dataset using each buffer distance and data version combination.

Overall, the 250m buffer performs the best across all data versions and model types (except for the null model, which is the naive model used for comparison). Therefore 250m is the selected buffer distance for further studies.

### *2.2.5. Models Training*

Applying a 250m buffer and the selected feature list, the complete dataset is filtered based on the dist_to_road feature to form the filtered dataset used for training. The rows with any null values are dropped. A random 80-20 split is conducted to split the filtered dataset into training and testing datasets. Standardization is conducted to the numerical features, and the same standardization scaler is saved and reused in testing.

We choose five commonly-used supervised learning algorithms for training. Logistic regression (LR) is the most basic classification machine learning algorithm that is simple to implement

*Table 2: Model Performance Summary - Aggregated Testing*

| | Data Version V1 | | | Data Version V2 | | | Data Version V3 | | | Data Version V4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Accuracy | F1 | AUC | Accuracy | F1 | AUC | Accuracy | F1 | AUC | Accuracy | F1 | AUC |
| LR | 67.2% | 0.675 | 0.661 | 68.9% | 0.689 | 0.665 | 74.0% | 0.734 | 0.702 | 73.9% | 0.733 | 0.699 |
| CART | 68.8% | 0.686 | 0.662 | 70.9% | 0.710 | 0.691 | 70.9% | 0.712 | 0.702 | 74.8% | 0.745 | 0.718 |
| RF | 69.9% | 0.683 | 0.645 | 72.1% | 0.716 | 0.686 | 75.4% | 0.745 | 0.708 | 74.9% | 0.740 | 0.703 |
| GB | 69.8% | 0.680 | 0.641 | 70.1% | 0.692 | 0.656 | 75.7% | 0.752 | 0.721 | 75.9% | 0.750 | 0.713 |
| HGB | 69.4% | 0.686 | 0.654 | 71.0% | 0.708 | 0.681 | 74.7% | 0.738 | 0.701 | 75.6% | 0.747 | 0.709 |
| Null | 62.7% | 0.483 | 0.500 | 63.4% | 0.409 | 0.500 | 63.5% | 0.494 | 0.500 | 63.6% | 0.495 | 0.500 |

and computationally efficient. The decision tree (CART) model is chosen because of its ease of handling both categorical and quantitative values. Random forest (RF) is selected because it works well with non-linear data and has a low risk of overfitting. Gradient boosting trees (GB) can be more accurate than random forests but may be more prone to overfitting and are slow to train. The histogram gradient boosting (HGB) algorithm accelerates the training speed of gradient boosting by binning the continuous input variables.

When training each model, 3-fold cross validation is conducted with halving grid search to speed up the cross validation process and conclude on the best model hyperparameters. After the optimal hyperparameters are decided from the cross validation, they are used to retrain the model with the entirety of the training data. When a road segment has no buildings nearby, the power line type is assigned as overhead, assuming that in remote areas, overhead power lines are installed because of the cost-benefit.

### *2.2.6. Classification Aggregation*

After the classification labels are produced at a building level, the final step is to aggregate the labels back to the road segment points. At this step, we experimented with different aggregation methods and decided to aggregate by mode, where we assign the popular prediction label at the building level within the 250m buffer to the center road segment. We have also conducted an inverse distance weighted aggregation method, where we first calculate the distance between each building and road point pair and aggregate the label using the inverse of distance as the weight for the weighted average. The two methods did not show a significant difference. Therefore, aggregation by mode is chosen because of its lower computational cost.

## 3. MODEL PERFORMANCE

The models are tested both at the individual building level and the aggregate road segment level, while model performance is evaluated at the aggregated road segment level. Aggregated testing accuracy and other metrics are summarised in table 2.

The Null model is simply assigning the majority of power line types to all road segments. It is producing different aggregated testing accuracies for each data version because the rows with null values are dropped, resulting in a slightly different number of rows for different data versions. From the aggregated testing results, we can see that for data versions with fewer features, simpler models such as Logistic Regression and CART decision trees have better performance. Gradient Boosting and Random Forest outperformed the simpler models for data versions with more features.

Looking at the mean feature importance extracted from the random forest model, the building year-built is the most significant contributing feature. The second most important feature is the building property value. Wind speed, time zone, building total area, political stance, number of floors, and occupancy are also among the top contributing factors to power line type classification.

When comparing across data versions, we can see that the data version V3 and V4 have similar aggregated testing accuracy. However, both of the data versions have at least 12 features. Among the 12 features, hazard related variables such as wind speed and flood depths will need to be interpolated at building levels, and urban/rural classification variable requires spatial processing using the

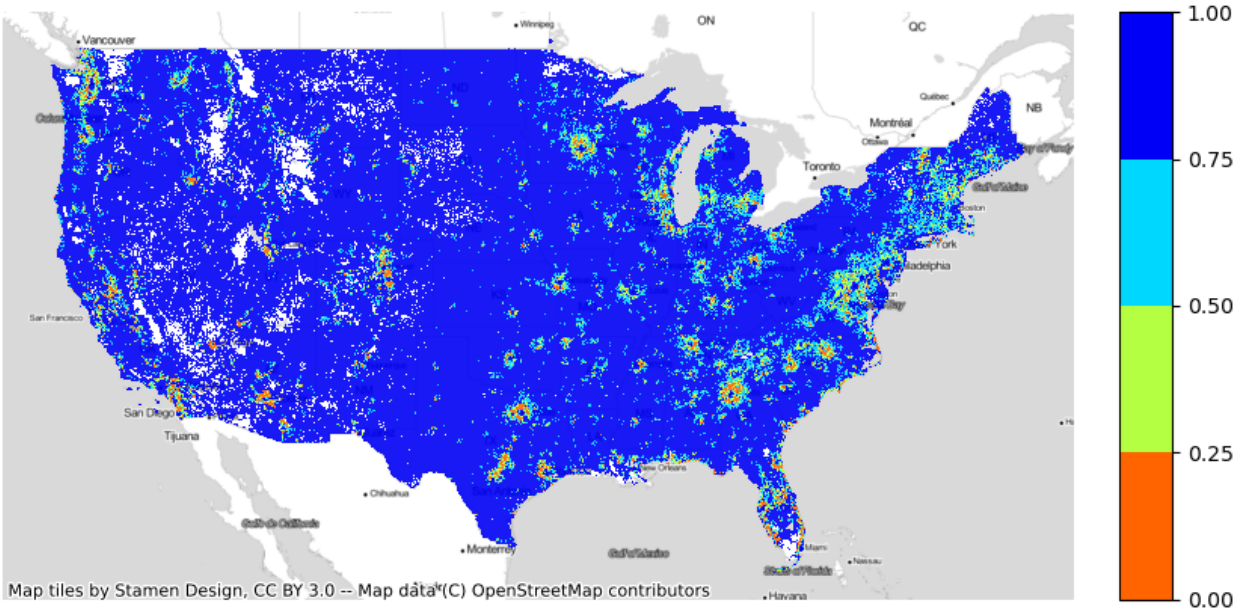


*Figure 3: Percentage of overhead power lines within each 5km x 5km grid cell of the U.S.*

complex shapefile. If data versions V3 or V4 are used for the classification of the power line types of all the road segments in the entire US, obtaining the input dataset will be computationally expensive. In comparison, compiling the input dataset using data version V2 is much easier, considering all the features are in one single building characteristics dataset. Its aggregated testing accuracy is also only 5% lower than data version V3. Therefore, data version V2 with its best-performing random forest model, is used for the power line type classification of the entire US.

## 4. NATIONWIDE POWER DISTRIBUTION NETWORK TYPE

After the random forest model for data version V2 is used for running the entire US, the percentage of overhead power lines aggregated at 5km x 5km grid cell level is shown in Figure 3. The hot spots are located in most major cities, suggesting that most of the power distribution networks are underground, and the model is indeed capable of capturing such a phenomenon.

We also zoomed into several major cities in the U.S. to look at the power distribution network classification in finer resolution. Figure 4 shows the percentage of overhead power lines within each 250m x 250m grid cell for six of the major US cities. For New York City (NYC), Chicago, Miami, Los Angeles (LA), and San Francisco (SF), we can clearly observe the hot spot for underground power lines at the city center. For downtown Houston, however, the model failed to capture the prevalence of underground power lines. In fact, we have observed that for smaller city downtowns, it is more difficult for the model to correctly identify the downtown underground power lines.

## 5. CONCLUSION

This study has presented the approach of using machine learning for the classification of power distribution networks. The model was trained and tested on the ground truth dataset consisting of around 60,000 road segments and their surrounding buildings. The best model achieved an overall accuracy of 75.9%. However, considering the model complexity and data scalability, we choose to use the model with only 5 features for the classification task of the entire US: building's year-built, total area, property value, number of floors, and occupancy type, and still maintained an accuracy of 72.1%. Overall, this study presents a novel approach to classifying power line types, which enables power outage simulation on a nationwide

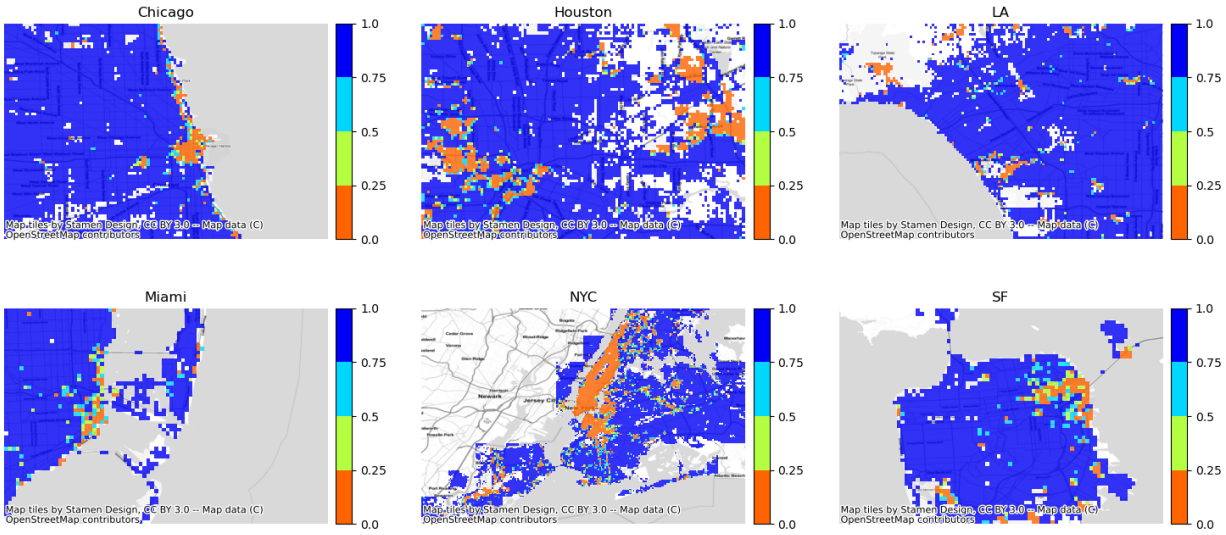


*Figure 4: Percentage of overhead power lines within each 250m x 250m grid cell at six major cities*

scale.

The proposed model has limitations. One aspect that could influence the network type but was not considered in this study is the demographics of different regions, such as household median income, education, and employment status. The model performance can be further improved by using more accurate data or by combining with more advanced deep learning approaches such as image-based object detection using street view images.

## REFERENCES

Chakalian, P., Kurtz, L., and Hondula, D. (2019). "After the lights go out: Household resilience to electrical grid failure following hurricane irma." *Natural Hazards Review*, 20.

Miura, Y., Qureshi, H., Ryoo, C., Dinenis, P. C., Li, J., Mandli, K. T., Deodatis, G., Bienstock, D., Lazrus, H., and Morss, R. (2021). "A methodological framework for determining an optimal coastal protection strategy against storm surges and sea level rise." *Natural Hazards*, 107(2).

Nateghi, R., Guikema, S. D., Wu, Y. G., and Bruss, C. B. (2016). "Critical assessment of the foundations of power transmission and distribution reliability metrics and standards." *Risk Analysis*, 36(1).

Prates, R. M., Cruz, R., Marotta, A. P., Ramos, R. P., Simas Filho, E. F., and Cardoso, J. S. (2019). "Insulator visual non-conformity detection in overhead power distribution lines using deep learning." *Computers  Electrical Engineering*, 78, 343–355.

Roussel, J.-R., Achim, A., and Auty, D. (2021). "Classification of high-voltage power line structures in low density als data acquired over broad non-urban areas." *PeerJ Computer Science*, 7, e672.

U.S. Census Bureau (2017). "U.S. Census Bureau TIGER/Line shapefiles. data retrieved from `https://catalog.data.gov/dataset/tiger-line-shapefile-2017-2010-nation-u-s-2010-census-urban-area-national`.

U.S. Geological Survey, National Geospatial Technical Operations Center (2022). "USGS National Transportation dataset (NTD). data retrieved from `https://data.usgs.gov/datacatalog/data/USGS:ad3d631d-f51f-4b6a-91a3-e617d6a58b4e`.

Waseem, M. and Manshadi, S. (2020). "Electricity grid resilience amid various natural disasters: Challenges and solutions." *The Electricity Journal*, 33, 106864.

Zhai, C., Chen, T., White, A., and Guikema, S. (2020). "Power outage prediction for natural hazards using synthetic power distribution systems." *Reliability Engineering  System Safety*, 208, 107348.

Zhou, Y., Pahwa, A., and Yang, S.-S. (2006). "Modeling weather-related failures of overhead distribution lines." *Power Systems, IEEE Transactions on*, 21, 1683 – 1690.