5-2016

Classifying Nominal Voltage of Electric Power Transmission Lines Using Remotely-Sensed Data

Erik Herman Schmidt
University of Tennessee - Knoxville, eschmid8@vols.utk.edu

To the Graduate Council:

I am submitting herewith a thesis written by Erik Herman Schmidt entitled "Classifying Nominal Voltage of Electric Power Transmission Lines Using Remotely-Sensed Data." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Geography.

Budhenrda L. Bhaduri, Major Professor

We have read this thesis and recommend its acceptance:

Nicholas Nagle, Bruce Ralston

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Classifying Nominal Voltage of Electric Power Transmission Lines Using Remotely-Sensed Data

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Erik Herman Schmidt
May 2016

Acknowledgements

Thank you to my committee members, Drs. Bhaduri, Nagle, and Ralston.
And thank you to my amazing wife, Sophie.

Abstract

Geospatial data of national infrastructure are a valuable resource for visualization, analysis, and modeling. Building these foundation-level geospatial infrastructure data sets presents numerous challenges. Among those challenges is that of acquiring non-visible attribution of particular infrastructure entities for which there is no viable tabular source. In the case of electric power transmission lines, these data are difficult to acquire, particularly nation-wide. The route, or geometry of transmission lines can be determined from aerial imagery, but nominal voltage, a fundamental requirement for analysis and modeling, is not readily apparent. However, inferences can be made about the nominal voltage based on visual characteristics, or predictors. This study develops a methodology to extract predictors from high-resolution aerial imagery and test the efficacy of those predictors for classifying the nominal voltage of transmission lines using a supervised classifier.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Critical Infrastructure and Geospatial Data Sets

As its designation implies, critical infrastructure is a selection of national infrastructure that is crucial to a country. On February 12, 2013, the White House released Presidential Policy Directive 21 (PPD-21), which outlined sixteen sectors of critical infrastructure. Communications, Emergency Services, Food and Agriculture, Transportation Services, and Energy are a few of the critical infrastructure sectors identified in this directive (1). According to PPD-21, the infrastructure within these sectors provides services that are critical for the nation to function (1); therefore, increasing and preserving their stability is of great national importance (1).

From a national perspective, the importance of critical infrastructure data arguably equals that of the physical infrastructure components themselves. In particular, geospatial data of critical infrastructure are a valuable resource for many communities for the purposes of visualization, analysis, and modeling. Of the three strategic imperatives articulated by PPD-21, the second states the priority to "Enable effective information exchange by identifying baseline data..." (1). This imperative not only underscores the importance of critical infrastructure data but also identifies the need for data to be shared effectively. The directive also charges federal agencies "to map geospatially, image, analyze, and sort critical infrastructure…" (1). So while identifying and sharing data is an overall priority for critical infrastructure security and resilience, geospatial data is of particular importance. An existing program that carries out this directive is the Homeland

Security Infrastructure Program (HSIP), which combines high-quality geospatial infrastructure data provided by all levels of government and the private sector for use by many communities, particularly Homeland Defense, Homeland Security, and National Preparedness - Protection, Prevention, Mitigation, Response, and Recovery (NP-PPMR&R)(2). The HSIP databases, which encompass foundation-level geospatial data sets from all critical infrastructure sectors, are used by the aforementioned communities for many applications, including "planning, situational awareness, threat and impact analysis (natural or man-made), modeling emergencies, protection of borders, and decision making during response and recovery operations" (2). In short, these geospatial data offer considerable utility to many communities for numerous applications. This study stems from an effort to create one of these geospatial data sets for the HSIP community.

The Problem – Nominal Voltage Data

Building these geospatial infrastructure data sets presents numerous challenges. Among those challenges is that of acquiring non-visible attribution of particular infrastructure components for which there is no publically available tabular source. This issue presents itself in many of the data sets within the energy sector. In the case of electric power transmission, the route, or geometry, of an overhead transmission line can be identified from high-resolution aerial imagery. However, nominal voltage data is not widely available in public sources.

Nominal voltage is a valuable datum in a transmission data set. Voltage is the difference in potential energy between two points on a circuit. The nominal voltage of a

transmission line is the voltage at which the line is designed to operate. These data are an essential input for power flow modeling and simulation. For HSIP users, particularly the NP-PPMR&R community, nominal voltage data provide emergency planners and responders with a gauge of a particular transmission line's relative importance as an infrastructure asset. However, voltage data for a particular transmission line may not be available in public sources. Herein lies the problem that this study will address: how can this gap in nominal voltage data be filled for a geospatial data set of transmission lines?

Electric Power Transmission Background

Transmission lines transmit electricity over long distances from electric power generators to substations near major load centers at nominal voltages above 69 kilovolts (kV) (3; 4). The majority of transmission lines in the U.S. are three-phase alternating current (AC)(3). Transmission lines above 69kV and up to 230kV are referred to as high voltage lines, while 345kV, 500kV, and 765kV are considered extra-high-voltage lines (EHV)(3). A single-circuit AC transmission line is comprised of three conductors, or three bundled conductors—one conductor, or bundle, per phase of AC (3). For the purposes of this study, transmission circuits will be referred to as transmission lines.

Though the electric power transmission industry includes many nominal voltages, transmission owners tend to use a small selection of voltages to provide their territory with power (4; 3). The reasons for this are both historical and economic. Generally, when peak energy demand exceeds four times the peak demand at the time the current highest transmission voltage was introduced, the utility will introduce new transmission lines with twice the voltage of the existing lines (4). Likewise, adjacent utilities generally

service their territory with similar voltages since it is more economical to build interconnections between transmission infrastructure without a transformer (4). Therefore, the voltage portfolios of neighboring utilities are likely to be similar, whereas utilities in different parts of the country may use voltages portfolios that are quite different from each other.

Inferences can be made about the nominal voltage of a transmission line based on its visual characteristics. Steel towers, for example, typically support very high voltage lines, while wood poles typically support lower voltage lines (3). Likewise, right-of-ways (ROW) are typically wider and support structures are usually taller when accommodating higher voltage transmission lines (3). What if characteristics such as these could be quantified as predictors of nominal voltage? What if those predictors could be extracted from high-resolution aerial imagery and used for a supervised classification?

Objective

Develop a methodology to estimate the nominal voltage of electric power transmission lines from high-resolution aerial imagery using supervised classification.

Research Questions

1. What predictors of nominal voltage can be extracted from aerial imagery and how can they be quantified?
2. How effective are these predictors in a supervised classification and how reliable is the classification?

Chapter 2

Previous Work

Introduction

Geospatial data sets of transmission lines, including voltage data, exist in various forms. However, existing sources of the transmission line data have significant limitations. Among the most notable communities in possession of these data are commercial vendors, utilities and electric power companies, and Volunteered Geographic Information (VGI) projects. The data sets within these communities vary considerably, despite representing the same physical infrastructure. More specifically, the accessibility, shareability, and quality of these existing data sets differs. In an attempt to justify the value of creating a new transmission line data set using the methodology outlined in this study, each of the aforementioned data sets will be assessed in terms of accessibility, shareability, and data quality.

Commercially Licensed Data Sets

Arguably, the best geospatial data of U.S. transmission lines are owned and maintained by commercial vendors. Three of the forerunners in this sector are Platts, Ventyx, and MAPSearch.com. Platts, a division of McGraw Hill Financial, provides expertise, as well as spatial and market data pertaining to energy, petrochemicals, metals, and agriculture, to customers from numerous industries, utilities, and government agencies (8). Ventyx, an ABB Company, provides similar services and information, with an additional emphasis on enterprise software for use by various industries (9). The PennWell Corporation's MAPSearch.com offers nationwide GIS data sets of various

energy infrastructure, including transmission lines (10), but very few details were found about this data set.

The accessibility of their respective transmission data sets varies slightly between Platts and Ventyx, while the use restrictions attached to their products are similar. As commercial products, access to these data sets requires that they be purchased. By purchasing these copyrighted data, users agree to abide by their commercial licenses. However, the manner in which these data sets are made available to users differs. Whereas Platts data can be ordered as a standalone GIS layer (11), Ventyx offers transmission data within their commercial software application called EV Energy Map—a component of Velocity Suite, their conglomeration of software and data for the energy industry (12). However, the Ventyx data set can also be accessed apart from their software. The U.S. Energy Information Administration (EIA) hosts a selection of Ventyx transmission line data for public viewing outside EV Energy Map via their U.S. Energy Mapping System and other interactive web maps (13). Due to copyright restrictions, however, these transmission data do not include voltage, are not available for download, nor are they viewable at scales larger than 1:4,622,324 (13).

In terms of data quality, the geospatial transmission data sets provided by Platts and Ventyx are perhaps the best available for the U.S., but each has notable issues. As a testament to the overall quality of the Ventyx data set, the Homeland Security Geospatial Concept of Operations (GeoCONOPS), an initiative to improve coordination of geospatial activities among all levels of government, the private sector, academia, and the public (14), documents the Ventyx data set as an authoritative geospatial data source and

cites that it has been incorporated into the HSIP database (15). The Platts data include transmission lines for North America with voltages between 110kV and 765kV in addition to some lower voltage lines (16). The attribution of the data set is extensive, effectively capturing data about primary, secondary, and tertiary line ownership, the number of circuits represented by the line, the type of line, the status, the substations on either end of the line, positional reliability, and voltage (16). With regard to horizontal positional accuracy, the data comply with, and in some areas exceed, National Map Accuracy Standards for a 1:250,000 map scale (16). However, the data lineage, or provenance, which describes the source, acquisition method, compilation, and any derivations of the information (17; 18), is not provided (16). The lineage of a data set is perhaps the most important element in data quality, particularly when assessing how appropriate the data are for a given application (17). Without any metadata to exposit the history and development of the data set, its quality is uncertain. This issue is compounded by the absence of a source field in the attribution that could offer insight about the provenance of particular features (16), especially those that are not immediately apparent based on appearance, such as voltage. It must be noted, however, that this assessment is based on the metadata available on the Platts website.

The Ventyx data set appears to suffer from the same quality issue, though to a lesser degree. Although not publicly available on the Ventyx website, selections of EV Energy Map transmission line metadata, as well as selections of the data, are available in the public domain as a part of the Bureau of Land Management's Rapid Ecological Assessment of the Middle Rockies and Northwestern Plains ecoregions (19; 20; 21).

According to the metadata, published in June, 2013, the data are a part of EV Energy Map and include transmission lines typically 115kV and above that have been clipped to the boundary of ecoregions (20; 21). The metadata do not explicitly name Ventyx as the data supplier. Nevertheless, as the source for the non-ESRI attributes, the metadata cite Global Energy Decisions—a provider of software and information that was acquired by Ventyx in June, 2007 (20; 21; 22). The attribution fields are nearly identical to the Platts data set, including voltage, but also include a source field to preserve the origin of each feature (20; 21). However, the lineage section of the metadata is empty, apart from citing EV Energy Maps (20; 21). This void could exist due to a failure to preserve the complete lineage information from the Ventyx data set when the data were copied. Regardless, an examination of the attribute data does not dismiss uncertainty surrounding the provenance. The source field for these data sets includes a combination of designations such as “Aerial Imagery”, “USGS Digital Line Graph”, “Regional Maps”, “Company Maps”, “Company Digital Data”, “Holding Company Maps”, and “EV Research” (20; 21). For a given feature, the field typically includes two or three sources, the first of which is one of the first six sources listed above (20; 21); the final source provided in this field is “EV Research”, with a few exceptions (20; 21). Six of these designations point to reasonable sources for transmission line geometry, but the last and most common designation, “EV Research”, which must refer to the origin of the tabular data values, describes a method, rather than a particular source. Given that the source field cites an unspecified method as the origin of the tabular data associated with particular features, users are left to make a judgment about the reliability of the data based on the credentials

of the originator, rather than an authoritative source. Thus, the provenance, and therefore, the quality of the tabular data—particularly voltage—is uncertain.

In addition to issues with accessibility and provenance, commercially licensed data sets have restrictive terms of use. As licensed data, these products cannot be shared by the product's consumer, or licensee. Distributing licensed data to any person, organization, or agency other than the licensee would be a violation of the terms of use. While not necessarily a significant restriction for some uses of transmission line data, this stipulation can be notably constraining if data are intended for distribution to a wide community of numerous agencies and organizations.

Utility and Electric Power Company Data Sets

Utilities and electric power companies that own transmission lines also have geospatial transmission data, but access to these data sets is limited. The Tennessee Valley Authority releases information regarding their ongoing and future transmission projects which includes a PDF map of the proposed route and the surrounding electric power infrastructure (23). However, the geospatial data displayed in these maps is not made available to the public. Similarly, Xcel Energy Inc. and American Electric Power release information pertaining to select ongoing transmission projects but access to the GIS data is not open to the public (24; 25). Other utility companies disclose limited information about current projects and instead encourage interested parties to contact them for additional information (26). Furthermore, utility companies may not own the geospatial data they use, as suggested by project maps released by Texas Entergy, Inc., a

part of the Entergy Corporation, which cite the Ventyx Velocity Suite as the data source (27).

In contrast to the previous examples of very limited data access, the utilities in Minnesota combined their geospatial data into a unified data set as a part of the Minnesota Electric Transmission Project, a collaborative effort between the Minnesota Department of Commerce and the Minnesota Geospatial Management Office (28). A statewide PDF map of the data set is publicly available. Accessing the GIS data, however, requires a formal request (28). Furthermore, the use constraints, as articulated in the metadata, prevent users who have been granted access to the GIS data from transmitting or sharing the data with a third party (28). These restrictions are in place due to security concerns (28). Whether this a common reason for utilities and electric power companies to restrict access to their data or an exception, cannot be determined. Regardless of the reason or reasons, accessibility remains a primary issue for utility and electric company-owned data.

Volunteered Geographic Information

Electric power infrastructure has not been passed over by the VGI community, particularly the OpenStreetMap (OSM) project. In sharp contrast to the data sets discussed previously, OSM allows unrestricted access to its database, which contains data of numerous geospatial phenomena worldwide, including roads, buildings, airports, railway stations, and more (29). These data are collected, contributed, and edited by thousands of ordinary volunteers either working independently or in project-oriented

groups (29). OSM contributors collect data in the field using GPS and out-of-copyright local maps, or remotely by tracing aerial imagery, with a particular reliance on local knowledge (30; 31). With regard to electric power infrastructure, OSM contains locations and basic descriptions of generation plants, substations, transmission and distribution lines, towers, and poles (32). Contributors have the option to record voltage values for transmission and distribution lines (32), and based on the ITO web map, attribution of 230kV lines and above appear extensive (33).

While OSM data can be freely accessed and used by the public, it bears a restriction regarding its use. The Creative Commons Open Database License (ODbL), under which OSM data are licensed, allows users to distribute, create, and adapt the data, provided users properly attribute OSM and share derivations of the data under the same open license if they are used publicly (29). The ODbL preserves the intentions of the contributors to distribute their data openly and prevents entrepreneurial parties from repackaging the data under an alternate license.

The most pertinent issue with OSM data and VGI in general, pertains to data quality. The issue stems from the volunteered nature of the data. Data are collected by ordinary people with varying amounts of local knowledge and experience with data collection (29). As a guide to help contributors determine line voltages, the OSM Wiki provides the following simple, yet problematic metric: “the length of the isolator (separating wires from tower) is 1 meter per 100 000 Volt” (34). Aside from potential error involved in estimating these measurements from the ground, the metric stated above is not valid. The rating of an insulator depends on its design, material, and configuration

(3), not its length directly. If a contributor possessed this knowledge, their estimates could be quite accurate. However, without a way to assess the qualifications of contributors and no quality control (29), verifying the accuracy of volunteered voltage data becomes a necessary requirement for establishing reliability of the data. In the absence of local knowledge or access to an alternate, authoritative source, verifying voltage data remains an issue for the OSM data set.

Summary

While geospatial transmission data sets with voltage are already available, they have limitations in the areas of data accessibility, shareability, and quality. Access to and shareability of commercial and utility data sets is limited—the former, due to licenses, the latter, security concerns. The provenance of voltage data included in commercial products can be unclear, while the reliability of these data, in the case of VGI, is entirely unknown. A data set created using the method outlined in this study has advantages over the aforementioned data sets in the key areas previously addressed. In particular, this study contributes a clearly defined, quantifiable, and repeatable methodology to estimate nominal voltage which can be used in the creation of a more accessible and shareable transmission line data set.

Chapter 3

Data and Methodology

Study Area

This study focused on a selection of transmission lines in the state of Minnesota. There are five commonly used transmission voltages in Minnesota—namely, 69kV, 115kV, 161kV, 230kV, and 345kV. By limiting the geographic extent to a state, the complexity of the classification should be less than if a wider area was considered, since a wider region might include additional voltages and therefore, additional classes. The reason for selecting this study area was pragmatic, since it was the only area where authoritative transmission data could be acquired.

Training Data

The training data used in this study was created from data provided by the Minnesota Geospatial Management Office, mentioned previously, which included a shapefile of transmission lines for the state of Minnesota. From this shapefile of 3,528 transmission line features, a selection of 56 transmission lines was taken for training purposes.

The selection of training features was made based on voltage, spatial distribution, and positional accuracy. Of the total features in the Minnesota data set, 3,457 features belong to one of the five voltages used in this study: 69kV, 115kV, 161kV, 230kV, and 345kV. A map of all transmission lines in this data set—symbolized by voltage—can be seen in Figure 8 (see Appendix). Other voltages in the shapefile include 34kV, 42kV,

138kV, 250kV, 400kV, and 500kV. These voltages, represented by 71 features, were excluded from the study due to their rarity—many corresponded to fewer than three transmission circuits. At least eight features were selected as training lines for each voltage in an effort to adequately represent each voltage class. Table 1 shows the number of original line features in the Minnesota data set by voltage class. The number of original line features is misleading. For example, while there were 2,218 69kV line features in the original data set, only 503 circuits could be identified. As many as 29 original line features could comprise a single transmission circuit. Therefore, the original features were dissolved based on their unique circuit designations. As a result of this step, 1,216 original features could not be successfully tied to a circuit and were therefore excluded from subsequent steps. Training lines were selected from the remaining 886 circuits.

Table 1. Original Minnesota transmission line data, circuits, training lines by voltage

Voltage (kilovolts)	***Original Line Features***	***Original Circuits***	***Sample Lines***
69	2,218	503	13
115	785	287	14
161	152	30	10
230	176	42	11
345	126	24	8
Total	3,457	886	56

In an effort to adequately sample a variety of transmission lines across the entire state, training features were also selected based on their location. Rather than examining lines clustered in a small region of the state, a dispersed collection of lines was selected. By selecting lines distributed across the state, the effect of differing geographies on transmission line characteristics could be better represented. Likewise, this approach attempted to account for the effect of varying transmission line construction practices of multiple transmission owners. For example, one transmission owner may prefer using wood supports for 115kV lines while another favors steel supports for the same voltage. If the selection of lines were limited to small area with a single transmission owner, this variation between owners would be missing from the training data set.

Likewise, the training selection was limited to features with good positional accuracy. Upon further examination of the original data set, numerous line features were ruled out because their corresponding transmission lines could not be identified from aerial imagery. Only features located close to their physical phenomena were practical candidates for the training data set.

Drawing from existing applications of supervised classification in remote sensing, there are at least two approaches that could have been undertaken in this study. Conceptually, the transmission circuits in this study could be pictured as the equivalent of objects in an object-based classification. However, if each circuit is thought of as a single object, then the total number of objects in the training data would be small, comprised of fifty six lines. Furthermore, if an entire line, which may be many miles long, is associated with only one value for each predictor, any variation in the line's characteristics along its

entire length would be lost. However, if a conceptual comparison is drawn between this study and pixel-based classification, then a single point along a given transmission line could be conceptualized as a pixel in an image. Just as an object in an image is represented by numerous pixels, each with slightly different values, so a transmission line can be represented with numerous observations by taking multiple measurements at different locations along the line. By using this cluster method, the size of the sample data could be expanded and each class can be more proportionally represented. Likewise, this method was expected to more accurately capture variation along each line.

A random sample of seven locations, hereafter referred to as sample points, were selected along each line feature. For a given line, the location of these points was selected by generating seven random digits between zero and the total length of the line. The locations associated with these measurements was then identified via linear referencing. The sample points were then moved to the closest support structure. These data were stored as point features in the File Geodatabase. Predictor measurements, detailed below, were made in ESRI's ArcMap environment at each of these locations—three-hundred ninety-two in total. A map of the sample lines and sample points used in this study can be seen in Figure 1.

Predictors

Overview

Predictor measurements were collected via image interpretation methods of 0.3 and 1-meter resolution aerial imagery. Acquiring adequate imagery in the exact locations

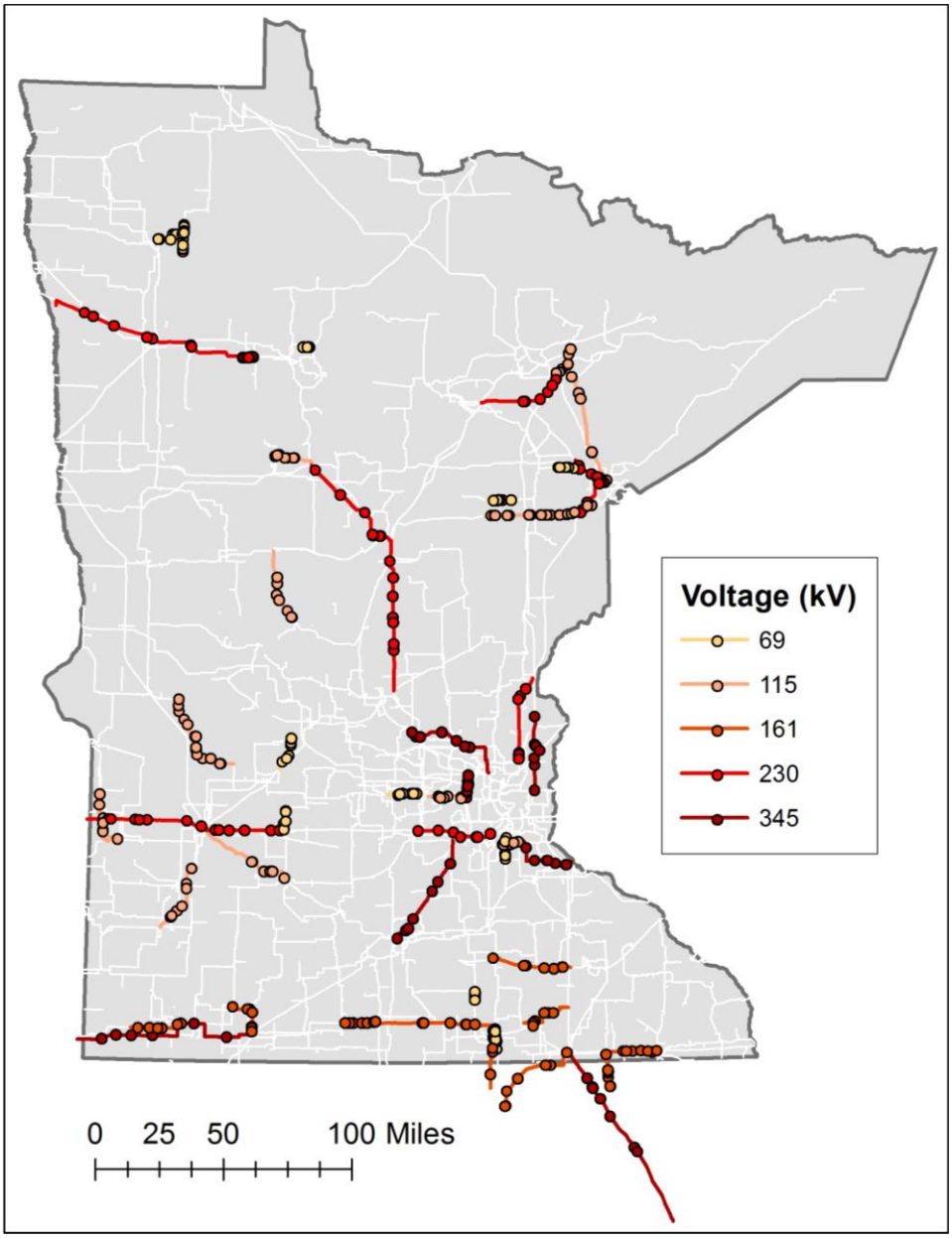

Figure 1. Map of Minnesota transmission lines and sample points used in this study

needed proved quite time-intensive. Most predictor measurements could be made from tiled basemap imagery, with the exception of support height. These measurements required imagery with a precise date and time of when the image was captured. Finding, acquiring, and managing these images proved slow. Once acquired, these images, along with the basemap imagery, were used to interpret and measure predictors.

Support Height

Generally, a higher clearance between the ground, or other obstacles, and the conductors is required the greater the line voltage, which necessitates taller supports (3). The distance between the base and the top of the support varies greatly depending on the support type and orientation of the conductors. A single pole with vertically oriented conductors sacrifices greater height for a smaller footprint. An H-frame line of the same voltage would not be as tall, but would have a wider footprint since the conductors are oriented horizontally. In an attempt to minimize this variation, support height was measured from the center of the base of the support to the cross-section or post insulator with the lowest conductor, rather than the top of the structure.

These measurements were taken from 1-meter resolution aerial imagery using mensuration tools available in ArcMap. Assuming an image has precise metadata on the date and time it was taken, the mensuration tools can be used to estimate the height of a structure by measuring the length of its shadow, assuming flat topography and a vertical structure.

Acquisition of aerial images with precise metadata proved quite challenging, and the process entailed avoidable setbacks. Images were acquired from DigitalGlobe via the

EnhancedView Web Hosting Service. In the first attempt, batches of images were selected, mosaicked, and downloaded. During collection of support height measurements, the date of one image, which depicted a snow-covered landscape, was identified as falling within the summer months. Upon review, the measurements collected from these images had no discernable relationship to voltage, which prompted closer examination of the images and metadata. It was then discovered that metadata records had been duplicated as a part of the mosaicking process. To avoid this issue in the final attempt, images were selected and downloaded individually, prior to measurement collection.

Support Span

The support span predictor was measured as the average distance between the support and the closest supports on the circuit. This predictor was expected to be more indicative of support height than voltage and therefore, a poor predictor. It was, however, easier to measure than the support height. However, the span is affected by the local topography (3), so this predictor was suspected to be a weaker input for the classifier.

Phase Spacing

The phase spacing predictor was measured as the average distance between the each phase of the circuit. Higher voltage lines require greater spacing between phases to ensure insulation standards are met (4). Collecting measurements of structures with horizontally-aligned phases was straightforward; structures with vertically-aligned phases was not feasible.

Right of Way Width

The right of way width predictor was measured as the distance between the tree line on either side of and perpendicular to the circuit. In the event that the circuit shared a right of way with another circuit, the distance from the support to the nearest tree line was measured and then doubled. Typically, utilities are required to maintain wider cuts through trees and vegetation to accommodate higher voltage lines and reduce the risk of circuit failure due to falling trees (4). An obvious constraint of this predictor is that it cannot be collected in areas without trees or tall vegetation.

Insulator Type

The insulator type predictor was measured as a binary variable. Insulators were identified as being either string insulators, which are suspended beneath a cross-section or arm of the support, or post insulators, which are attached above a cross-section or mounted perpendicularly to a single pole support. Post insulators are typically used on lower voltage transmission lines with lighter conductors (3), making them a suspected predictor of voltage. Generally, insulator type was determined by first identifying the support type or by examining the support shadow.

Support Type

The support type predictor was comprised of qualitative variables that describe the type of structure supporting the circuit. As an example, a support type could be a single pole, a double pole, otherwise known as an H-frame, or a tower. This variable was suspected to be an important prerequisite to a more accurate classification, not necessarily as a predictor of voltage directly, but rather, as a predictor of sub-classes. As

an example, a 115kV line supported by a single pole is likely to have a more narrow right-of-way than a 115kV line supported by a double pole, or H-frame, support since the former aligns the conductors vertically and therefore, occupies less horizontal space than the latter (3). Depending on this predictor, subsequent predictors may exhibit differing distributions within the same class, which was examined prior to classification.

While initially collected as three binary variables, corresponding to single pole, H-frame, and pylon, these variables did not perform well in the classification. In attempt to improve the importance of this predictor, as determined by the goodness measure, the H-frame and pylon variables were combined, and then merged with the single pole variable to create a single, binary predictor of support type.

Support Material

The support material predictor was comprised of a qualitative, binary variable that describes the construction material of the support: wood, or metal/concrete. Wood-pole supports are generally most economical for lines up to 230kV (3), while the mechanical strength of concrete and especially metal make them optimal for higher voltage lines (3). Distinguishing between concrete and metal poles was found to be infeasible, so these variables were merged.

Multi-circuit

The multi-circuit predictor was measured as a binary variable. Some transmission supports are designed to accommodate more than one transmission circuit. Based on observations of transmission lines in the MN data set, these multi-circuit supports appeared to be more prevalent among some voltage classes.

Bundled Conductors

The bundled conductors predictor was also measured as a qualitative, binary variable that indicated whether each phase of AC is carried by more than one conductor. Bundled conductors are commonly used on EHV transmission lines to reduce corona discharge, whereby the air surrounding an energized conductor is ionized, resulting in power loss (4). Ideally, this variable would be measured on a continuous scale, since the number of bundled conductors tends to increase with higher voltage (3), but making this distinction was infeasible using 0.3-meter imagery so a binary measurement was used.

Classification

Classification trees are a form of decision tree, and like all classifiers, are used to predict categorical outcomes based on observations with known outcomes, or classes (35). The classification is accomplished in part through a processes of segmenting, or splitting the data into groups based on rules (35). If observations were plotted on an x-y plane, classification trees successively divide this plane into regions in an attempt to minimize the heterogeneity of classes within regions based on the training data set (35). Conceptually, this method resembles a tree where each rule splits the data into branches, at the end of which are terminal nodes, or leaves (35). An example of a tree produced in this study can be seen in Figure 2.

The classification tree method was used due to its simplicity and suitability for qualitative variables. Although decision-tree based methods do not perform as well as other classifiers, they can be interpreted relatively easily (35). Likewise, classification

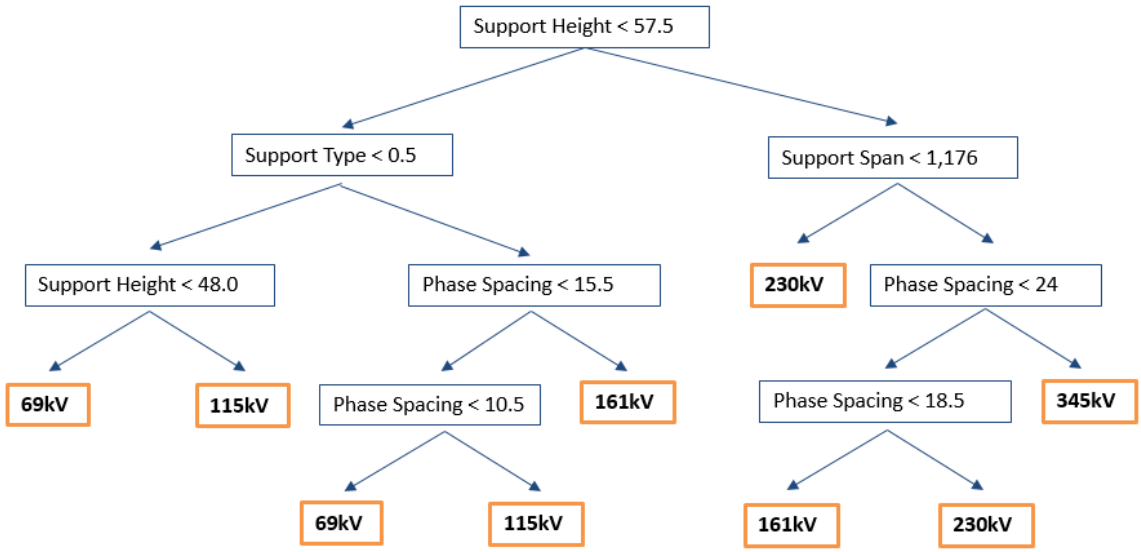

Figure 2. A classification tree produced using this method

trees are well-suited for qualitative predictors (35), of which there were many in this study. As for their weaker performance compared to other classifiers, for this foundational study, this was an acceptable trade-off for clearer interpretability. This study is primarily concerned with a new application of classification methods, and therefore, the framework of the overall voltage classification method must be affirmed before additional, more complex classifiers are introduced. Once this foundational methodology has been tested, future work could involve the use of alternative classifiers to improve the predictive accuracy.

Classification trees in this study were created in R using the rpart package. The rpart module uses a recursive partitioning method that repeatedly partitions observations using maximal impurity reduction criteria, or rules that attempt to split the data such that

the impurity, or heterogeneity, of nodes are minimized (36). Prior to a split, all predictors are ranked according to their impurity reduction criteria—or goodness measure—and the best is used to split the observations, after which the process is repeated (36). Predictors that are used for these splits are referred to as primary variables (36). In the event that an observation is missing a value for a primary variable, a surrogate variable is used (36). The surrogate variable and split are determined using the partitioning method again to best separate observations into the nodes decided by the original primary variable split (36).

In order to avoid overfitting a model to the training data, the tree must be pruned, or restrictions placed on the number of partitions, or splits. Overfitting occurs when the model conforms too closely to the training data—indicated by numerous splits—to be useful when applied to other data. Trees in this study were pruned by selecting the complexity parameter—and therefore the number splits—related to the lowest cross-validation error. Rpart calculates a complexity parameter (cp), or cost associated with adding additional partitions to the tree (36). A lower cp results in more splits and increased risk of overfitting the training data, so rpart also performs 10-fold cross-validation for each additional split (36). Selecting the complexity parameter tied to the lowest cross-validation error is an attempt to parse the tree to a size that adequately classifies the observations without overfitting the training set.

To train and test the tree, the sample point data set was divided into a training data set and a test, or validation data set. This partition was executed based on the sample lines, rather than the sample points. In this way, sample points from the same sample line

were not used to both train and test the classification; rather, all sample points belonging to each individual sample line were either used for training or testing purposes.

The validation set approach was used to evaluate the performance of the classification tree when applied to new data. At the sample line level, the data were divided into two equal halves with 196 sample points from twenty-eight sample lines in each half. The first half was used to train the classification tree, while the other half was set aside to test the performance of the tree. Since the performance of the classification depends greatly on which samples were used in training and which we set aside for testing, this splitting processes was performed randomly in thirty iterations.

The predictive accuracy and Kappa-Coefficient were calculated and recorded for each validation set iteration. Predictive accuracy was measured simply as the number of true positives divided by the total number of observations in the test set, represented as a percentage. The Kappa-Coefficient is a measure of classification performance that, unlike predictive accuracy, accounts for chance accuracy, and is calculated as

$$K = \frac{\Pr(correct\ classification) - \Pr(chance\ classification)}{1 - \Pr(chance\ classification)}$$

where the probability of correct classification is the number of true positives divided by the total test observations (37). Chance classification is calculated by first dividing the product of observations belonging to class x and classified as x by the total number of observations squared, then adding this value for all classes (37). By this measure, a

Kappa-Coefficient of 0 would indicate the supervised classification was no better than a random classification and 1 would indicate a perfect classification.

Likewise, the user accuracy and producer accuracy were calculated for each class in each iteration. The user accuracy is the probability that the actual class of a given observation is *x* considering it was classified as x, and was calculated as the number of true positives for that class divided by the number of observations classified as that class (37). The producer accuracy is the probability that an observation belonging to class x will be classified as x, and was calculated as the number of true positives for that class divided by the number of observations belonging to that class (37).

As a part of post-processing, the final output of the classifier was aggregated to the line features. Each line was assigned to the class with the highest frequency based on the mode of its sample points. As an example, if four out of seven sample points associated with a single transmission line feature were predicted to be from the 161kV class while the remaining three, the 230kV class, the transmission line would be assigned to the 161kV class. In this way, outlier sample points that were misclassified can be smoothed over in the final cluster classification.

Chapter 4

Results

Exploring Predictors

Continuous Predictors

An examination of the predictors revealed a distinct signature, of varying strengths, with respect to voltage, especially among the continuous predictors. As expected, support height and phase spacing, exhibit a distinct, positive relationship to voltage. As shown in Table 2, the mean of these predictors increased with each consecutively higher voltage. The lowest mean in each of these variables is found in the lowest voltage, 69kV, while the highest mean was associated with the highest voltage. Exceeding expectations, the support span predictor exhibited an equally distinct signature with respect to voltage, as shown by the mean which increased in each consecutive class. To a degree, the right of way predictor shares this trend, but the signature was less distinct. While overall, the mean right of way was less in the lower classes and greater in the higher classes, the mean did not increase with each sequentially higher voltage. In particular, the mean right of way for the 161kV lines stood out as an outlier because it was greater than the 230kV class and nearly equal to the 345kV class. This was likely due to a disproportionately high number of missing right of way values in the 161kV class compared to other classes, as evident by the Nulls field. Therefore, among the continuous predictors, support height, support, span, and conductor spacing were found to exhibit noticeable, positive relationships to voltage, while a weaker trend was shown by the right of way width predictor.

Table 2. Summary statistics of continuous predictors by voltage

	69kV	***115kV***	***161kV***	***230kV***	***345kV***	***Total***
Sample Points	91	98	70	77	56	392
Support Height						
Min.	22.0	27.0	32.0	47.0	65.0	
Mean	37.6	46.9	57.1	69.5	83.5	
Std. Dev.	8.7	6.0	10.7	13.2	11.6	
Max.	70.0	64.0	88.0	110.0	126.0	
Nulls	0	0	3	0	0	3
Support Span						
Min.	225.0	209.0	300.0	581.0	474.0	
Mean	414.1	573.1	613.7	967.2	980.8	
Std. Dev.	103.0	134.3	197.4	245.6	141.4	
Max.	798.0	788.0	1000.0	1660.0	1264.0	
Nulls	0	0	0	0	0	0
Conductor Spacing						
Min.	9.0	10.0	15.0	12.0	15.0	
Mean	10.1	12.5	16.2	21.0	25.1	
Std. Dev.	0.7	1.6	1.1	3.6	4.4	
Max.	11.0	16.0	20.0	29.0	34.0	
Nulls	80	24	30	7	26	167
Right of Way Width						
Min.	40.0	65.0	90.0	90.0	110.0	
Mean	73.8	90.8	143.3	125.1	150.6	
Std. Dev.	18.7	12.2	55.1	18.1	69.9	
Max	130.0	120.0	200.0	160.0	320.0	
Nulls	65	60	67	47	48	287

While overall, a given continuous predictor might exhibit a distinct, general relationship to voltage, some classes displayed stronger signatures than others. As an example, Figure 3 shows a frequency plot of support height wherein almost every class features a dominant prominence along the x-axis. The 161kV prominence, however, is overshadowed by the 115kV and 230kV classes, which was a common characteristic of the 161kV class across all continuous predictors. The effect of this lack of dominant signature for this class was seen clearly in the classification performance.

Furthermore, the variation of continuous predictor values differed notably across voltages. The standard deviation rows in Table 2 provide a gauge of the variation within classes for any given predictor. Compared to the 69kV, 115kV, and 345kV classes, the 161kV and 230kV had much higher standard deviations with respect to support span,

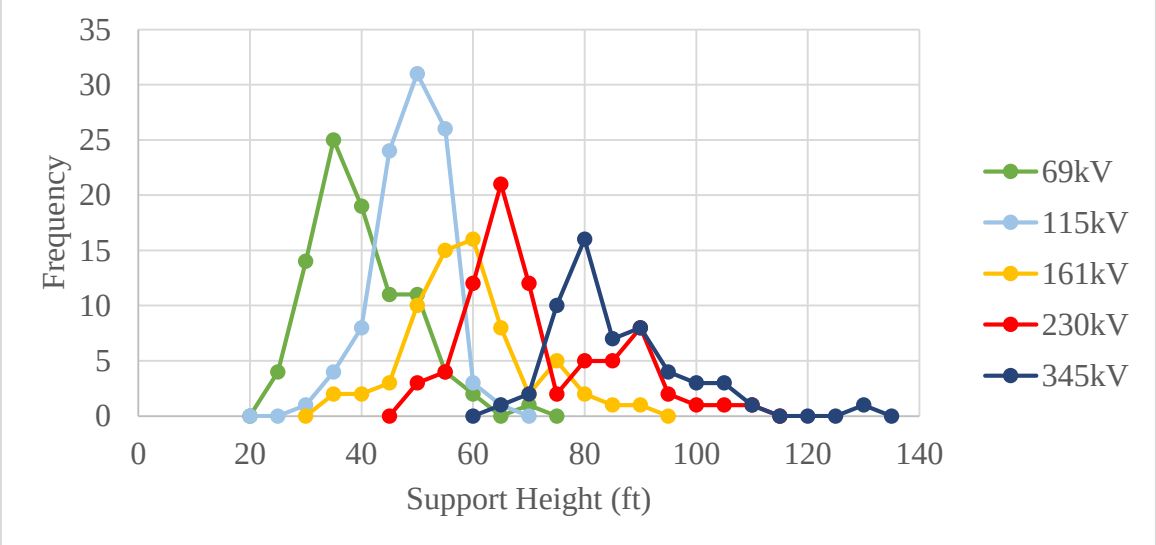

Figure 3. Frequency plot of support height by voltage

suggesting that these latter classes have more variation than the former classes. To a lesser degree, this trend was also seen in the support height of the 230kV class and in the conductor spacing of both 230kV and 345kV. With respect to support height, the differing shapes and widths of frequency curves in Figure 3 also illustrates this intra-class variation. The 115kV class has a narrow, near-normal curve while the 345kV class features a wide, positively-skewed shape. However, the conductor spacing predictor showed little variation within classes, excluding the classes mentioned. As noted above, industry standards are quite strict in regard to the spacing of conductors in order to minimize phase interference, which could account for this finding. Furthermore, the presence of any perceived variation within this predictor could be a result of inaccurate measurements during data collection.

In addition, the intra-class variation resulted in ranges that often overlapped with each other—some more so than others. The minimum and maximum rows in Table 2 showcase a wide range of values within any given class. The support span of the 69kV class, for example, varied from 225 to 798ft. while the 115kV class, from 209 to 788ft.—a considerable overlap, despite having means of 414 and 573ft., respectively. This overlap can also be seen in Figure 3. The range of support height of any given class overlapped with the range of every other class, which presented a challenge for a tree-based classification used in this study. The variation within and overlapping among classes is also evidenced in Figure 4, which plots support height against support span. The scatterplot displays relatively tight, uniform clusters of 69kV, 115kV, and, to a lesser

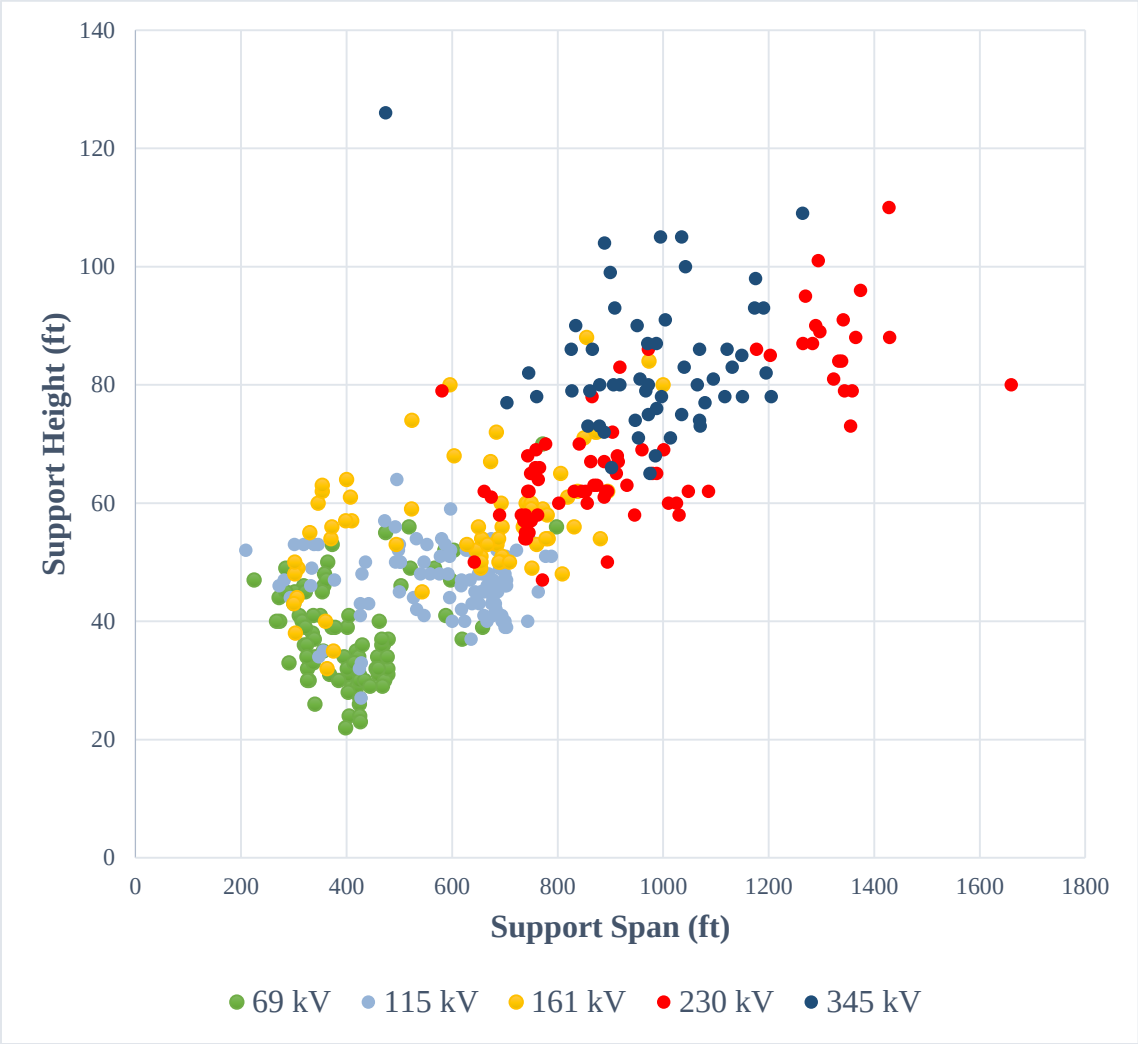

Figure 4. Support span and support height by voltage for sample points

extent, 345kV samples. The 161kV class, in contrast, is dispersed—lacking any strong clustering or dominance anywhere on graph. As is evident by the classification result, the absence of a strong signature in the 161kV across the predictors present negatively impacted the predictive accuracy of the classifier vis-à-vis the 161kV class.

Once again, conductor spacing was an exception to this trend of overlapping classes. Apart from the 161kV and 230kV classes, there was relatively little overlap among voltages, as seen in Table 2. Again, this was likely due to the nature of industry standards and the construction of transmission line supports.

Regarding variation in the 230kV class, Figure 4 provides further insight. Unlike the 161kV class, which lacked a strong signature in neither support height nor support span, the 230kV class exhibited two clear signatures, as evident by the two clusters on either side of the 345kV class. This class displayed intra-class variation in the form of two distinct distributions. The cause of this binary signature stemmed from ownership, and brought an important piece of geographic insight to this methodology. The point samples with the higher support height and support span were taken from 230kV transmission lines owned by two companies in the middle part of Minnesota. A map of transmission ownership, produced from the original Minnesota data set, can be seen in Figure 9 (see Appendix). These companies built their lines using pylon supports which, as seen in Figure 4, were taller and spaced further apart than other 230kV lines sampled in this study. The remaining samples were from 230kV lines located in the north, which are owned and operated by different companies than those in the middle of the state. The northern companies constructed their lines with wooden H-frame supports. Therefore,

while two transmission owners may utilize the same transmission voltage, the signatures of their respective lines may be vastly different. This finding suggests that consideration should be given to ownership, and therefore, geographic regions when training transmission voltage classifiers.

As expected, collecting some predictor measurements was infeasible for many observations. For example, right of way width could not be collected for 237 sample points, or approximately 73% of observations due to the absence of a tree line or tall vegetation. Likewise, 167 conductor spacing values, or 42% of observations could not be measured because the conductors were vertically aligned.

Binary Predictors

Signatures of voltage were less apparent among the binary predictors but visible, regardless. Most of the predictors displayed some distinctly lopsided classes, particularly for the lower and higher classes, as shown in Table 3. The support type of the 69kV class, for example, were mostly single poles, while the 115kV class was mostly H-frames or pylons. Likewise, the 230kV class was made up of exclusively H-frames or pylons. Similarly, bundled conductors were most commonly found in the 345kV class, rarely in the 230kV class, and never in the 69kV and 115kV classes. The 69kV class was made up of mostly wooden supports, while the 345kV class was mostly metal or concrete. Signatures in the middle classes were less distinct. For example, the 161kV class fails to clearly stand out in any predictor. In the support type variable, the class was nearly split evenly. In the insulator type and multi-circuit variables, it resembled lower classes. In

Table 3. Summary of binary predictors by voltage

	69kV	***115kV***	***161kV***	***230kV***	***345kV***	***Total***
Sample Points	91	98	70	77	56	392
Insulator Type						
String	73	88	56	77	56	350
Post	18	10	14	0	0	42
Support Material						
Wood	66	62	29	36	11	204
Metal / Concrete	25	36	41	41	45	188
Support Type						
H-Frame / Pylon	12	74	40	77	42	245
Single Pole	79	24	30	0	14	147
Bundled Conductors						
Yes	0	0	11	6	44	61
No	91	98	59	71	12	331
Multi-circuit						
Yes	1	3	5	17	26	52
No	90	95	65	60	30	340

bundled conductors and support type variables, 161kV displayed patterns similar to higher classes.

Classification Results

Predictor Performance

As was apparent after exploring the individual predictors, some predictors performed better than others in the classification. For each iteration of the validation set approach, wherein the data were randomly divided and half were used to train a tree while the other used to test, the predictors used as primary variables in tree construction and their rank—determined by variable importance—was recorded. The results can be found in Table 4.

Out of all predictors used in this study, the continuous variables—with the exception of right of way—performed best in the classification. Conductor spacing,

Table 4. Predictor performance in tree construction

	Total Trees as Primary Split	***Mean Variable Importance Rank***
Conductor Spacing	29	3.55
Support Height	27	1.60
Support Span	24	1.60
Bundled Conductors	15	4.18
Support Type	14	4.10
Multi-circuit	1	6.12
Support Material	0	6.25
Insulator Type	0	6.71
Right of Way	0	Null

support height, and support span were used as a primary variable for at least one partition in 29, 27, and 24 out of 30 trees, respectively. These predictors also performed best in terms of their average variable importance rank. The rpart package calculated variable importance by taking the sum of the goodness measure of all partitions where the predictor was used as the primary variable and the product of goodness measures where it was used a surrogate and the adjusted agreement, after which rpart normalized these values to combine to 100 for each tree (36). For each tree, variable importance was sorted and ranked between 1 and 9, with one corresponding to the highest variable importance in that tree. The mean of these importance ranks is found in column two of Table 4. By this metric, support height and support span were most effective in the classification, followed by conductor spacing. On the other hand, right of way width performed the worst of any predictor. It was never used as a primary or surrogate variable, likely due to wide variation and numerous missing values.

Performance of the binary predictors was mixed. Bundled conductors and support type were used as primary splits in 15 and 14 trees out of 30, respectively. The only other binary predictor used as a primary variable was multi-circuit, and only once. The remaining variables were never used for a primary split, but they served as surrogate variables, as indicated by their mean variable importance rank. Based on these two metrics, bundled conductors and support type performed best among the binary predictors.

Predictive accuracy

The overall performance of this method was fair, as seen in Figure 5. The mean predictive accuracy of thirty iterations was 62.1%. As expected from the validation set approach, the results varied greatly depending on how the samples were randomly divided into testing and training sets, which is shown by the wide range of predictive accuracies. The worst tree had a predictive accuracy of only 45.4% and the best tree, 70.9%. However, the overall distribution of the predictive accuracies was slightly negatively skewed by the 45.4% outlier, as suggested by the shape of the green curve in Figure 5. The frequency plot shows that most of the trees produced predictive accuracies between 60% and 70%. So while some trees yielded accuracies below 50%, the majority correctly classified 60% or more observations. Unsurprisingly, the Kappa-coefficient showed a more conservative assessment of overall performance. Figure 6 shows a similar frequency plot of Kappa-Coefficient for the thirty iterations. The mean Kappa-Coefficient was 0.523, and most trees produced scored greater than 0.5. While not a high score, this method consistently yielded results well above a chance classification.

A granular look at the classification result provided a more insightful assessment of its strengths and weaknesses. Table 5 includes a confusion matrix with the results produced by the tree seen above in Figure 2. Matrix columns are classes predicted by the tree and rows are actual classes. In this example, forty-four out of forty-nine 69kV samples were correctly classified as such, and likewise, seventeen out of twenty eight 345kV samples were correctly labeled. The middle class, 161kV, performed very poorly; twenty out of forty-two samples—less than half of the observations—were correctly

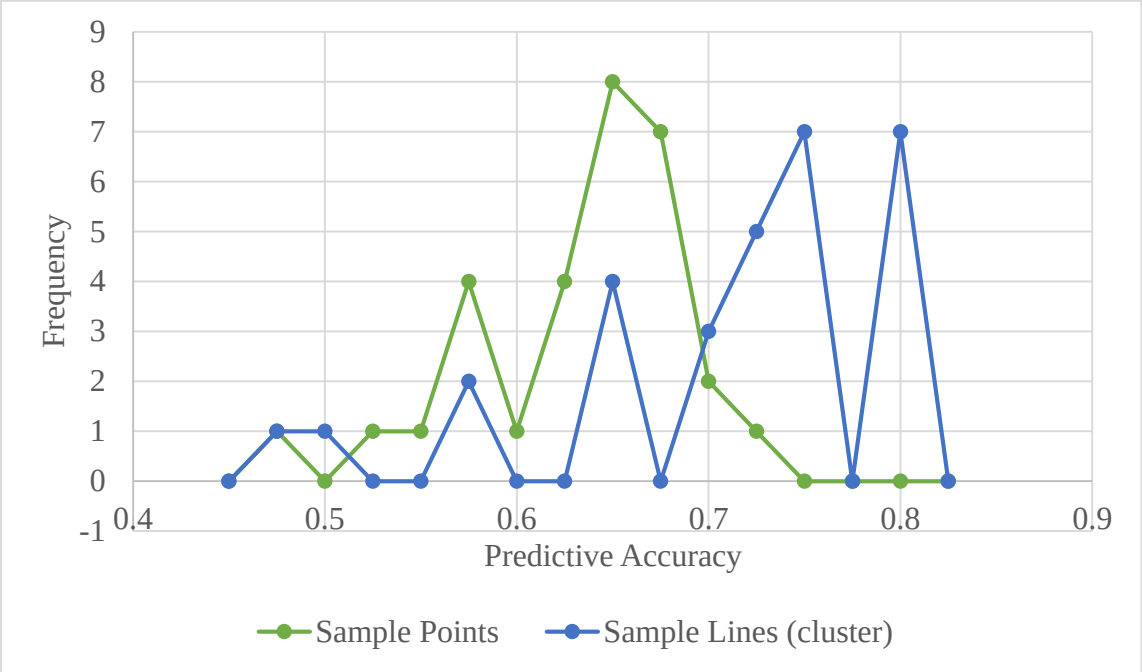

Figure 5. Frequency of predictive accuracy for thirty iterations

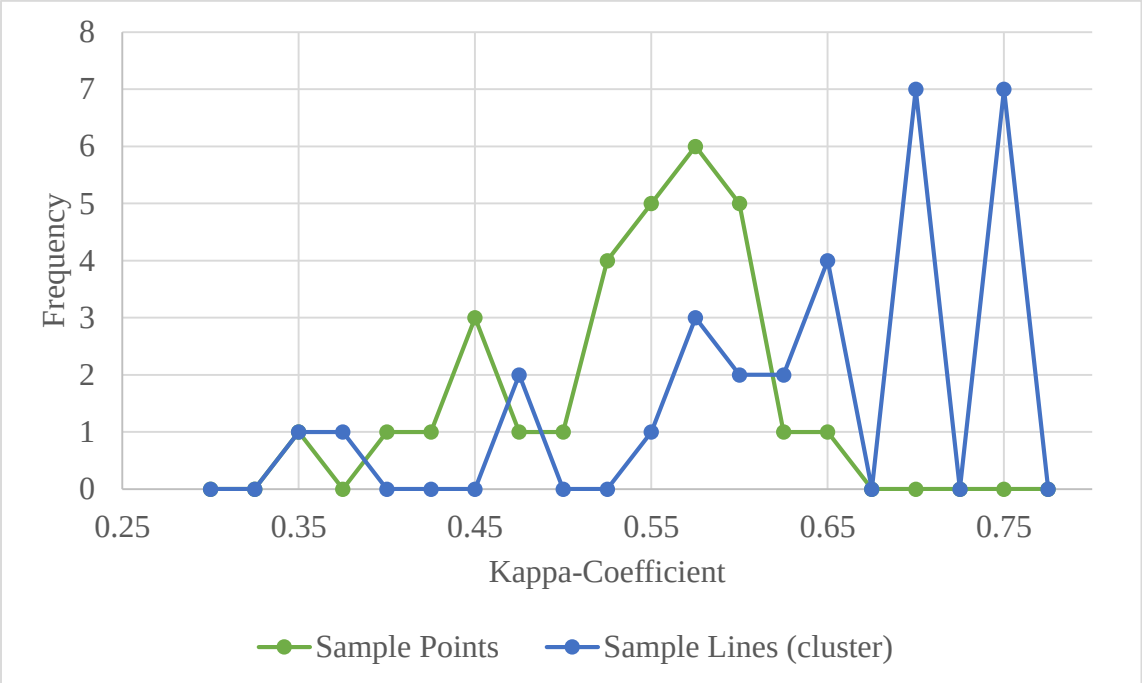

Figure 6. Frequency of Kappa-Coefficient for thirty iterations

classified, yielding a low producer accuracy. Furthermore, nineteen observations were incorrectly classified as 161kV, which gave it a low user accuracy as well. Given the weak signatures of the 161kV class relative to other classes, as seen above, this result was unsurprising. This result was not limited to this example, however. Table 6 shows the user accuracy and producer accuracy results for all iterations. As shown by the mean user and producer accuracies, this method consistently underperformed with regard to the 161kV class, relative to other classes.

The disparity between the model's ability to handle higher and lower voltage classes verses middle classes is seen clearly in Figure 7. This plot displays producer accuracy against user accuracy for all thirty iterations, symbolized by class. Wide variation was found in all classes with respect to these two metrics. However, most trees handled the classes reasonably well, as shown by the concentration in the upper right section of the graph. The 161kV class was the notable exception, as evident by the clustering of yellow points in the middle of the graph, distinctly separated from the majority of the other class scores. Given the weak signatures of the 161kV class

Table 5. A confusion matrix generated using this method

	Sample Points					***Sample Lines (Cluster)***				
	69kV	***115kV***	***161kV***	***230kV***	***345kV***	***69kV***	***115kV***	***161kV***	***230kV***	***345kV***
69kV	44	5	0	0	0	7	0	0	0	0
115kV	10	23	2	0	0	1	4	0	0	0
161kV	9	11	20	2	0	2	1	3	0	0
230kV	0	0	14	28	0	0	0	1	5	0
345kV	0	0	3	8	17	0	0	0	1	3

Table 6. Summary classification statistics for thirty iterations

	Min.	***Mean***	***Median***	***Std. Dev.***	***Max.***
Prediction Accuracy	45.4%	62.1%	64.3%	5.9%	70.9%
Kappa	0.332	0.523	0.536	0.069	0.632
Prediction Accuracy (cluster)	46.4%	70.1%	71.4%	8.4%	78.6%
Kappa (cluster)	0.346	0.623	0.639	0.103	0.729
User Accuracy					
69kV	0.396	0.682	0.679	0.146	0.962
115kV	0.311	0.635	0.639	0.152	0.960
161kV	0.000	0.483	0.471	0.179	1.000
230kV	0.000	0.705	0.747	0.197	0.931
345kV	0.282	0.665	0.697	0.203	1.000
Producer Accuracy					
69kV	0.408	0.740	0.806	0.173	1.000
115kV	0.286	0.620	0.639	0.176	0.898
161kV	0.000	0.397	0.405	0.176	0.714
230kV	0.000	0.684	0.750	0.230	1.000
345kV	0.095	0.746	0.779	0.163	0.952
No. of Primary Splits	3	5.9	6	1.5	8

compared to other classes, the classifier had difficulty distinguishing it from other classes, particularly 115kV and 230kV. A spatial perspective offered an explanation of and possible solution to this issue with the middle classes. A closer look at Figure 8 (see Appendix) revealed a significant pattern in the distribution of transmission voltages within the state. The 161kV class is only present in the southern part of the state, while the 230kV class is only found in the middle and northern regions. Likewise, the 345kV lines are predominantly located in the southern and middle regions, while the 115kV class is most commonly found in the middle and northern regions. In short, there is little overlap in the regions served by 161kV and 345kV lines and those served by 115kV and 230kV lines. The capacity of 161kV lines makes them a middle ground between 115kV and 230kV lines, so the overlapping signatures of this class with the others should have been expected since the classification was performed at the state level.

Given the ordinal structure of voltage classes, the trees in this study performed better than the overall predictive, producer, and user accuracy suggests. In practice, a classification would not be performed with so few observations in the training data set, as was the case for some trees included above. For this reason, trees with fewer than four sample lines, or 28 sample point observations, per class in the training data set were excluded from the following tabulation. The confusion matrices of the remaining seventeen trees were normalized by row, or actual class, and corresponding records across all matrices were summed, and then divided by seventeen to produce the average classification rate for each class, as shown in Table 7. The table shows that on average the vast majority of observations were classified within one position of their true class.

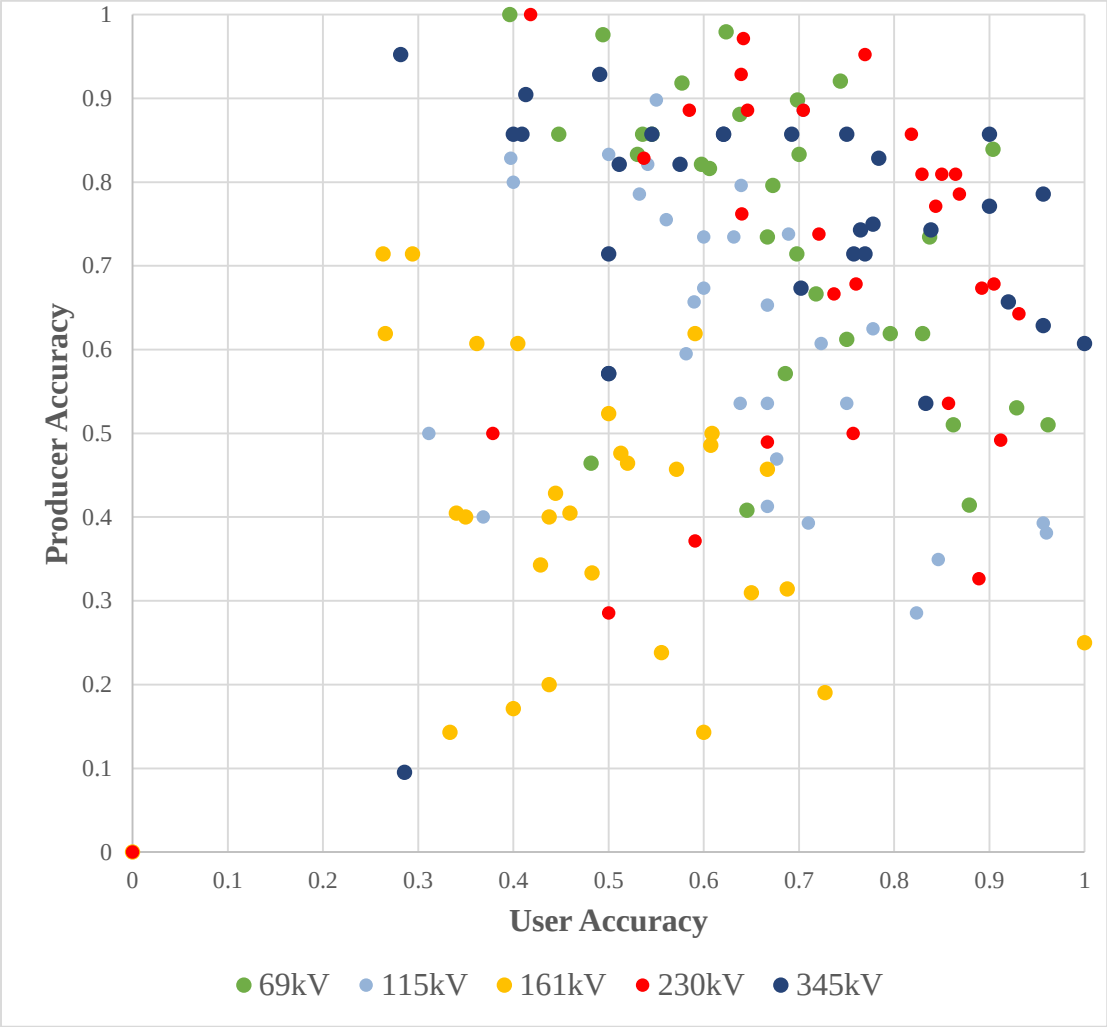

Figure 7. Producer and user accuracy results by voltage for thirty iterations

For example, an average of 0.966 observations belonging to the 69kV class, seen in the first row, were either classified as 69kV or one class higher, 115kV. On average, far fewer observations were classified as 161kV, fewer as 230kV, and even fewer, as 345kV. This trend can be seen in all classes, wherein the classification rate declines in positions further from the true class, with the exception of the 161kV class, which has a nearly even rate beyond the true class.

As expected, the final post-processing step in the clustering approach significantly improved the overall performance of the classification by smoothing out misclassifications. The smoothing effect of identifying the most frequently predicted class of the sample points as the final class of the corresponding sample line can be seen in the confusion matrix in Table 4. While the producer accuracy improved only slightly, the user accuracy increased greatly since many observations incorrectly classified as 161kV were smoothed out. Likewise, some noisy misclassifications in the 69kV, 115kV, and 345kV classes were smoothed over. Likewise, the effect of this post-processing step

Table 7. A confusion matrix of average classification rate by class across select iterations

	Sample Points					***Sample Lines (Cluster)***				
	69kV	***115kV***	***161kV***	***230kV***	***345kV***	***69kV***	***115kV***	***161kV***	***230kV***	***345kV***
69kV	0.791	0.175	0.023	0.009	0.001	0.858	0.142	0.000	0.000	0.000
115kV	0.264	0.595	0.136	0.005	0.000	0.222	0.657	0.121	0.000	0.000
161kV	0.162	0.192	0.393	0.116	0.136	0.157	0.201	0.422	0.102	0.118
230kV	0.000	0.006	0.139	0.668	0.186	0.000	0.000	0.088	0.776	0.137
345kV	0.000	0.002	0.040	0.179	0.779	0.000	0.000	0.029	0.088	0.882

on the average classification rate can be seen in the right-hand matrix in Table 7, wherein misclassifications by more than two class positions are entirely eliminated, and some, by just one class position. However, this smoothing effect also negatively affected the classification, as evident by the two 161kV lines misclassified as 69kV in Table 4—the equivalent of 14 misclassified sample points compared to the 9 misclassified sample points from the original classification. Figures 5 and 6 show the predictive accuracy and Kappa-Coefficient of each iteration's cluster classification, as compared to the original, sample point classification. The frequency curves of the cluster classification show significant improvement over the sample point classification. As indicated by the blue curve in Figure 5, most trees had a predictive accuracy greater than 70% after the clustering classification, and the best trees correctly classified 78.6% of observations. The mean predictive accuracy increased by 8.0% and the Kappa-coefficient by 0.099 as shown in Table 5. By taking the most frequently predicted voltage from the point samples as the most likely voltage of the transmission line, noisy predictions were smoothed out and the overall performance of the classification increased. Therefore, this simple step was a significant improvement over the original classification.

Chapter 5

Conclusions and Recommendations

This study developed and tested a foundational methodology to classify nominal transmission line voltage using measurements taken from aerial imagery. Using the methods outlined in this study, predictors of voltage were successfully measured from aerial imagery using imagery interpretation and mensuration techniques. Of the predictors identified and collected, the height of transmission supports and span between supports were found to be the most effective for classifying voltage, followed by the spacing between conductors, the type of support, and presence of bundled conductor phases.

The tree-based classification used in this study yielded fair classification performance. In the majority of iterations, the tree produced a predictive accuracy between 60% and 70% and Kappa-Coefficient greater than 0.5, up to 0.632. However, examination of the average classification rate showed that most misclassified observations were only one class higher or lower than the true class. Furthermore, both metrics mentioned above improved significantly in the clustered classification step. By identifying the most frequently predicted class for samples from a given line and labeling that line as such, noise in the classification was smoothed out and the overall performance improved. After this step, the majority of iterations reached a predictive accuracy greater than 70%, up to 78.6%., and a Kappa-Coefficient greater than 0.65, up to 0.729. Likewise, the clustering step eliminated misclassifications more than two class positions from the true class.

A granular look at the results of the classification with respect to individual classes revealed weaker performance among the middle classes, particularly the 161kV class. An exploration of the predictors showed weak signatures in the 161kV class compared to other classes. As a result, the trees developed in this study consistently yielded lower user and producer accuracies for that class, often mistakenly labeling observations as 161kV and classifying 161kV observations as other classes.

Further examination of the 161kV class with a spatial perspective revealed important insights. The 161kV class and its neighboring classes, 115kV and 230kV, were located in distinct, non-overlapping regions. This suggests that careful consideration should be given to delineating the geographic region within which the classification is applied. In so doing, the number of transmission voltages—and therefore, the number of target classes—in the classification is reduced. The class signatures, in turn, should be more pronounced. While this would requiring training more classifiers, the classification performance would likely improve.

A regional approach is also supported by examining signatures of transmission lines belonging to the same voltage class in different parts of Minnesota. By exploring the support height and support span of the 230kV class, two very different signatures were found—one corresponded to lines in the middle part of the state, the other, to lines in the north. Had lines from both of these regions not been sampled and used to train the classifier, the classification would have performed more poorly. Therefore, in future applications and expansions on this methodology, it should not be assumed that the

signature of a class as determined in one region necessarily applies to the same class in another region.

Although the results of this study suggest a regional approach to voltage classification to be advantageous, delineating appropriate regions presents a challenge. While some utilities may publish maps of their service territory—implying clearly defined boundaries—these boundaries are difficult to delineate. Transmission lines near these boundaries may be owned by another utility. Although neighboring utilities are more likely to use similar voltages, this should not be assumed. Future work on this subject should be devoted to addressing this challenge. One possibility for exploration would be to infer voltage boundaries from power plant data published by the Energy Information Administration, which includes the grid voltage at the point of interconnection for each power plant. Although not a comprehensive portfolio of nominal voltage, since it would not necessarily capture intermediate voltages between power plants, it may be used to generate a point cloud from which to delineate regions with common transmission voltage portfolios.

Another means of improving the voltage classification would be to generalize classes. Assuming companies in a given region use five or more transmission voltages, classifying voltage in that area may result in the same issues encountered in this study. However, if classes were merged into generalized classes, such as “below 100kV”, “115-161kV”, “230-287kV”, etc. the predictive accuracy would increase, at the expense of precision. This approach would generate a more conservative estimate of nominal voltage, while still offering a degree of distinction between transmission lines. The

average classification rate lends further support for this approach, given that most misclassifications were within on class of the true class. Future work could be devoted to determining the most appropriate voltage groupings.

Given the limited predictive capacity of tree-based methods, future study could implement and compare the results of other classifiers. As noted above, classifications trees were used in this study primarily for their advantage in interpretability over other methods. Because of the foundational nature of the methodology outlined in this study, this characteristic was essential. Since the groundwork has been established and the methodology produced a fair result, more robust classifiers, such as support vector machines and neural networks could be utilized to improve the predictive accuracy of this method.

One obstacle to the scalability of this method is the time-intensive nature of collecting predictor measurements for numerous sample points. To address this issue, future work could focus on reducing the number of sample points necessary to capture variation of a line while maintaining an acceptable accuracy threshold for the cluster classification. The methodology outlined in this study could be repeated but with a smaller collection of sample points from each line. The cluster classification could be repeated with fewer sample points and the overall performance assessed with each iteration to determine the optimal number of sample points.

Another opportunity to improve the scalability of this method would be to leverage non-imagery data sets to extract key predictors identified in this study. For

example, collecting support span and support height—the latter in particular—measurements was time-intensive using the methods in this study. Capturing support height required imagery with precise metadata about date and time, which can be difficult to acquire. However, both of these predictors could hypothetically be extracted from vertical aeronautical obstruction data collected by the Federal Aviation Administration, since transmission line supports fall under this designation.

The primary contribution of this study is a clearly documented lineage and repeatable methodology to estimate transmission voltage for the creation of foundational geospatial data sets. The output of this methodology has the potential to bring a data set with limited accessibility to a wider audience. For example, it could be useful to the open source community. Users could compare the output of a classifier created using this methodology to data from OSM and other VGI sources for validation purposes. Alternatively, trained participants could collect voltage data in the field from a selection of transmission lines, which could then be used to classify additional, less accessible lines. This could be particularly useful in other parts of the world where transmission data sets may have missing voltage data. On its own, this methodology may not be the best means of filling gaps in transmission data, nor was it intended to be. Rather, this methodology is intended as a supplement to other sources and means of collecting voltage data.

List of References

1. **Office of the Press Secretary.** Presidential Policy Directive -- Critical Infrastructure Security and Resilience. *The White House.* [Online] February 13, 2013. http://www.whitehouse.gov/the-press-office/2013/02/12/presidential-policy-directive-critical-infrastructure-security-and-resil.

2. **Homeland Infrastructure Foundation-Level Data Working Group.** Homeland Security Infrastructure Program (HSIP) Gold and Freedom. *Homeland Infrastructure Foundation-Level Data Working Group.* [Online] 2014. https://www.hifldwg.org/public/HSIP-Gold-Freedom-One-Pager-2013.pdf.

3. **Shoemaker, Thomas M and Mack, James E.** *The Lineman's and Cableman's Handbook.* 10th. New York : McGraw-Hill Companies, Inc., 2002.

4. **General Electric Company.** *Electric Utility Systems and Practices.* [ed.] Homer M Rustebakke. 4th. New York : John Wiley & Sons, Inc., 1983.

5. **Tennessee Valley Authority.** TVA Distributors. *Tennessee Valley Authority.* [Online] [Cited: November 5, 2014.] http://www.tva.com/power/pdf/tva_distributor_map.pdf.

6. —. TVA's Transmission System. *Tennessee Valley Authority.* [Online] [Cited: November 5, 2014.] http://www.tva.com/power/xmission.htm.

7. —. Right of way vegetation maintenence FAQ. *Tennessee Valley Authority.* [Online] [Cited: Nov 5, 2014.] http://www.tva.com/power/rightofway/faq.htm.

8. **Platts, McGraw Hill Financial.** About Platts. *Platts, McGraw Hill Financial.* [Online] 2014. [Cited: November 8, 2014.] http://www.platts.com/about.

9. **Ventyx, an ABB Company.** About Ventyx. *Ventyx, an ABB Company.* [Online] 2014. [Cited: November 8, 2014.] http://www.ventyx.com/en/about-ventyx.

10. **PennWell Corporation.** Electric Power GIS Data. *MAPSearch.* [Online] 2014. [Cited: November 8, 2014.] http://www.mapsearch.com/gis-asset-data/electric-power-gis-data.html.

11. **Platts, McGraw Hill Financial.** GIS Data. *Platts, McGraw Hill Financial.* [Online] 2014. [Cited: November 8, 2014.] http://www.platts.com/products/gis-data.

12. **Ventyx, an ABB Company.** Product Overview - Velocity Suite: Investment Grade Data Analysis. *Ventyx, an ABB Company.* [Online] [Cited: November 8, 2014.] http://www.ventyx.com/~/media/files/brochures/velocity-suite.ashx?download=1.

13. **U.S. Energy Information Administration.** Layers Information for Interactive State Maps. *EIA.* [Online] [Cited: November 1, 2014.] http://www.eia.gov/maps/layer_info-m.cfm.

14. **Department of Homeland Security.** About US. *GeoPlatform.* [Online] [Cited: November 1, 2014.] https://geoplatform.gov/geoconops/about-us.

15. —. Transmission Lines. *GeoPlatform.* [Online] [Cited: November 1, 2014.] https://www.geoplatform.gov/node/925.

16. **Platts.** Electric Transmission Lines. *Platts, McGraw Hill Financial.* [Online] April 9, 2009. [Cited: October 20, 2014.] http://www.platts.com/IM.Platts.Content/ProductsServices/Products/gismetadata/trans_ln.pdf.

17. **Guptill, Stephen C and Morrison, Joel L.** *Elements of Spatial Data Quality*. Oxford : Elsevier Science, 1995.

18. *A Survey of Data Provenance in e-Science.* **Simmhan, Y L, Pale, B and Gannon, D.** 3, 2005, Special Interest Group on Management of Data Record, Vol. 34, pp. 31-36.

19. **Bureau of Land Management.** Rapid Ecological Assessments. *Bureau of Land Management.* [Online] November 4, 2014. [Cited: November 15, 2014.] http://www.blm.gov/wo/st/en/prog/more/Landscape_Approach/reas/midrockies.html.

20. —. MIR_DV_C_Transmission Lines_ln.lyr. June 15, 2013.

21. —. NGP_DV_C_Transmission_Lines_ln.lyr. January 17, 2013.

22. **Businessweek.** Global Energy Decisions, LLC: Private Company Information. *Bloomberg Businessweek.* [Online] 2014. [Cited: November 15, 2014.] http://investing.businessweek.com/research/stocks/private/snapshot.asp?privcapId=3332679.

23. **Tennessee Valley Authority.** TVA: Transmission System Projects. *Tennessee Valley Authority.* [Online] 2014. [Cited: November 22, 2014.] http://www.tva.com/power/projects/index.htm.

24. **Xcel Energy.** Transmission Projects. *Xcel Energy.* [Online] 2014. [Cited: October 20, 2014.] http://www.xcelenergy.com/Safety_&_Operation/Transmission/Transmission_Projects.

25. **American Electric Power.** American Electric Power Transmission Projects. *AEP Transmission.* [Online] 2014. [Cited: October 20, 2014.] http://www.aeptransmission.com/.

26. **Georgia Power.** Building for the Future. *Georgia Power.* [Online] [Cited: October 20, 2014.] http://www.georgiapower.com/about-energy/delivering-energy/building-for-the-future/home.cshtml.

27. **Entergy Texas.** Ponderosa to Grimes Transmission Project. 2013.

28. **Minnesota Geospatial Information Office.** Metadata: Electric Transmission Lines and Substations, 60 Kilovolt and Greater, Minnesota, 2014. [Online] September 25, 2014. [Cited: October 20, 2014.] ftp://gdrs.dnr.state.mn.us/gdrs/data/pub/us_mn_state_mngeo/util_elec_trans/metadata/metadata.html.

29. **Kerski, Joseph J and Clark, Jill.** *The GIS Guide to Public Domain Data.* Redlands : ESRI Press, 2012.

30. **OpenStreetMap.** *OpenStreetMap.* [Online] [Cited: October 25, 2014.] http://www.openstreetmap.org/about.

31. *How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordinance Survey datasets.* **Haklay, Mordechai.** 2010, Environment and Planning B: Planning and Design, Vol. 37, pp. 682-703.

32. **OpenStreetMapWiki.** WikiProject Power networks. *OpenStreetMap Wiki.* [Online] 2014. [Cited: October 18, 2014.] http://wiki.openstreetmap.org/wiki/WikiProject_Power_networks.

33. **ITO World.** ITO Map - Electricity distribution. *ITO! Map.* [Online] 2014. [Cited: October 11, 2014.] http://www.itoworld.com/map/4.

34. **OpenStreetMap Wiki.** Key:voltage - OpenStreetMap Wiki. *OpenStreetMap Wiki.* [Online] 2014. [Cited: October 11, 2014.] http://wiki.openstreetmap.org/wiki/Key:voltage.

35. **James, Gareth, et al.** *An Introduction to Statistical Learning with Applications in R.* New York : Springer, 2013.

36. **Therneau, Terry M. and Atkinson, Elizabeth J.** *An Introduction to Recursive Partitioning Using the RPART Routines.* [Online] June 28, 2015. [Cited] March 1, 2016. https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

37. **Canty, Morton J.** *Image Analysis, Classification, and Change Detection in Remote Sensing: With Algorithms for ENVI/IDL.* 2nd. Boca Raton: CRC Press - Taylor & Francis Group, 2010.

Appendix

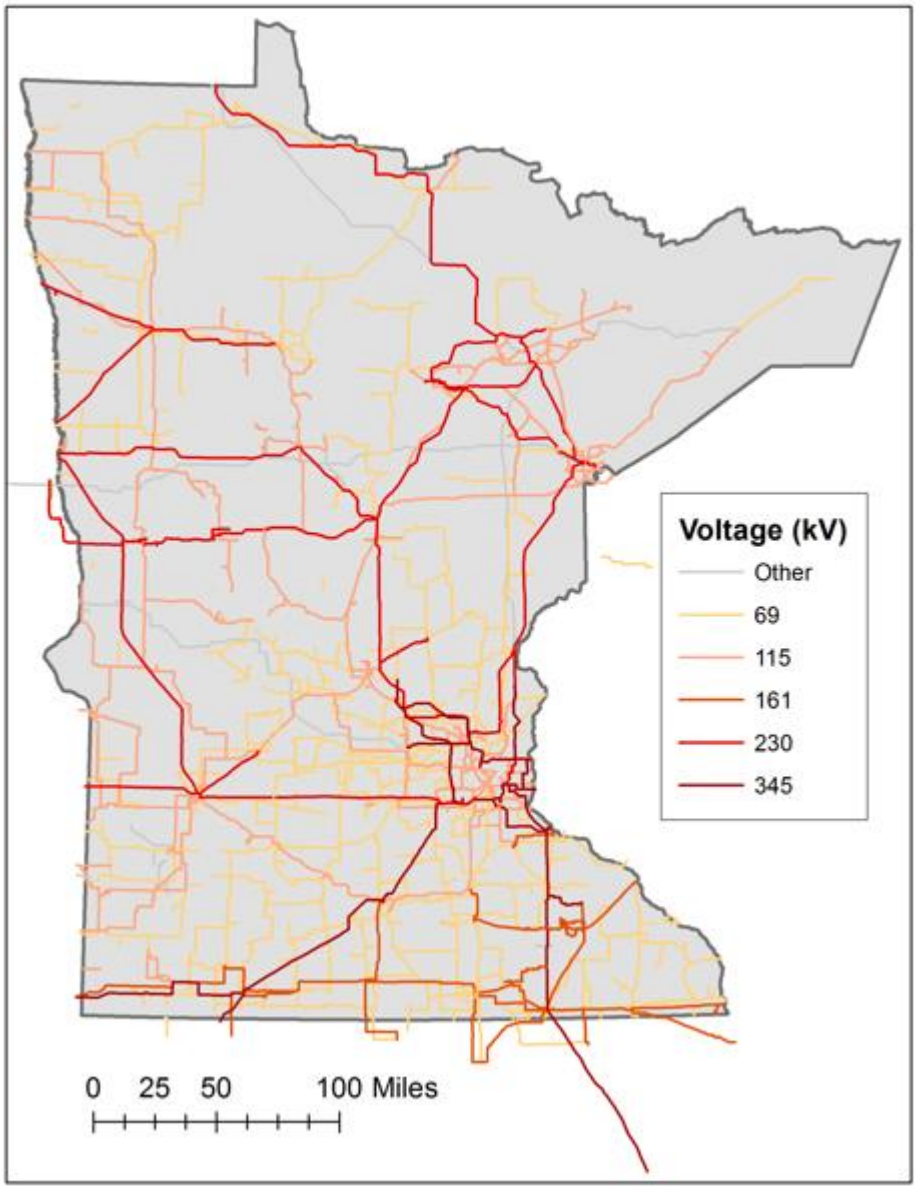

Figure 8. Map of transmission line voltage in Minnesota

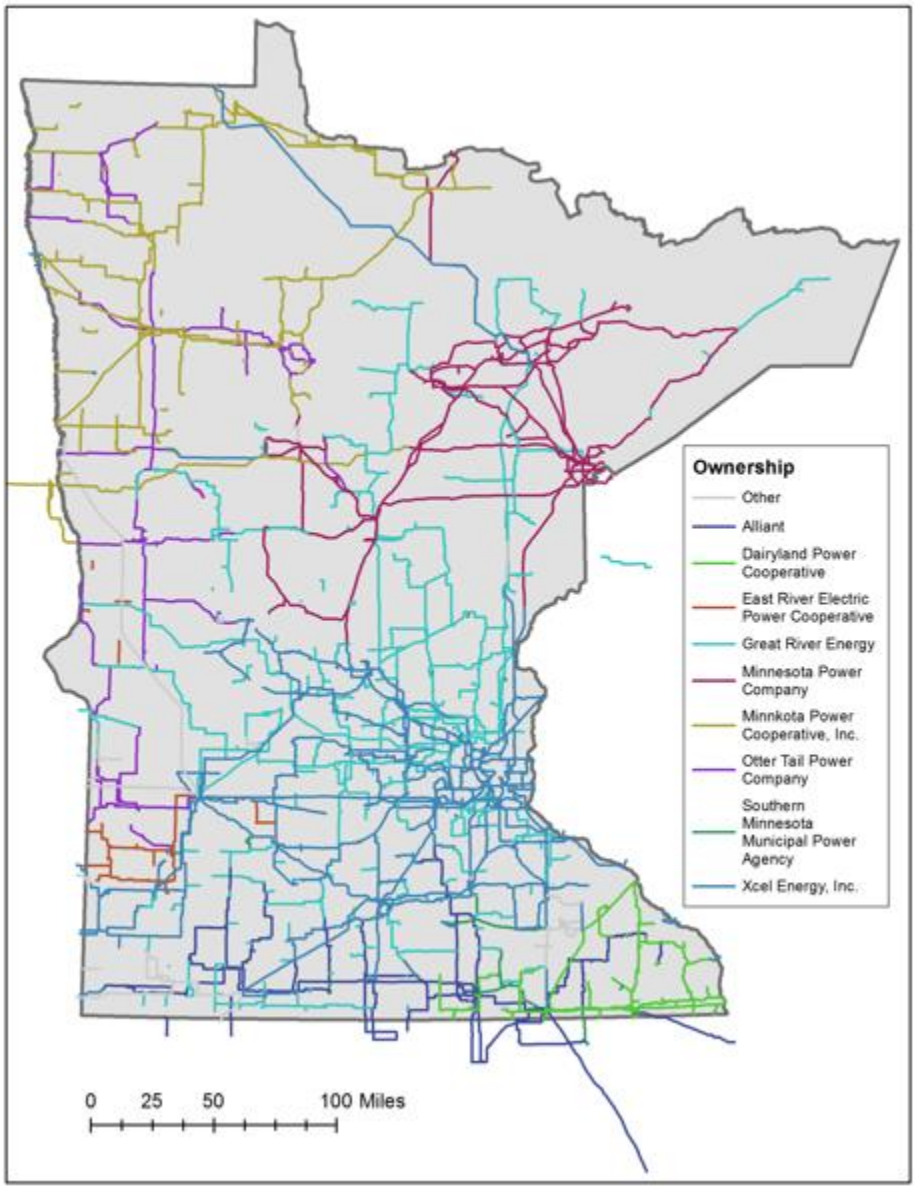

Figure 9. Map of transmission line ownership in Minnesota

Vita

Erik H. Schmidt was born on March 21, 1991 in San Diego, California. He was homeschooled by his parents till his high school graduation in 2009. After moving with his family to Oak Ridge, Tennessee, he enrolled in classes at Roane State Community College, where he served as a writing tutor and earned an Associate of Arts in 2011 before transferring to the University of Tennessee, Knoxville. Having received a Bachelor of Arts degree in geography in 2013 he stayed in Knoxville and began his graduate studies. He and his wife were married in their home church of Covenant Presbyterian Church in Oak Ridge in June of 2014. Under the advising of Dr. Budhu Bhaduri, mentoring of Mark Tuttle, encouragement of his parents, Kurt and Kristi Schmidt, love of his wife, Sophie, and by the grace of God, he wrote and defended his master's thesis before graduating in the spring of 2016.