Developing and evaluating a method to disaggregate legacy SSURGO soil maps

by

Jordan Taylor Watson

B.S., University of Dayton, 2016

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF ARTS

Department of Geography and Geospatial Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:
Major Professor
Dr. Arnaud Temme

Copyright

Abstract

For soil mapping, legacy (survey) data has its greatest potential in the disaggregation of complex soil mapping units. Disaggregation in this context is possible by relating different soils within a single map unit to other known characteristics of the landscape such as slope and curvature. For the United States soil map SSURGO, virtually all current soil map units have multiple soil types (known as “series”) described within them, indicating there is great potential for such disaggregation. In this study SSURGO map units are disaggregated using descriptions of soil map units and Digital Elevation Models (DEMs) for three locations in the central United States. The model presented disaggregates soil map units using information already within the soil description such as the curvature and slope in which a certain soil series is found. First, slope and curvature are calculated using National Elevation Dataset 10m resolution DEMs. Next, soils are assigned preferentially using fuzzy membership functions to the slope and curvature values for each map unit until all the possible soils are disaggregated. The area percentage of each new soil is calculated, and the model is adjusted until the SSURGO-stipulated percentage of each component is reached. The method worked best for county size areas and the newly created maps increased the number of map polygons by over 100%. The new maps were then compared to soil point data and showed similar accuracy to SSURGO and better accuracy than a similar digital soil mapping approach.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This project was funded as part of a larger USDA research project titled: “Disaggregating SSURGO soil maps across large areas using existing qualitative knowledge and modern data sources” (NR193A750023C008). This thesis was completed with support from my Major Professor, Arnaud Temme and my committee Shawn Hutchinson and Deanne Presley. Additionally, support was provided from the NRCS officials in Kansas (Jeffery Hellerich), Michigan (Matt Bromely, Jonathan DiazCruz), and Minnesota (Joseph Brennan).

Introduction

Soil maps are used for a variety of purposes ranging from scientific research, agricultural needs, and community planning. Additionally, the soil information within these maps is also crucial for measuring climate change indicators such as soil carbon storage (Causarano et al., 2008; Davidson & Lefebvre, 1993). Because of their importance across different fields, soil maps are available in a wide range of scales, resolutions, and for many geographic regions. The United Nations' Food and Agricultural Organization and Educational, Scientific and Cultural Organization (FAO/UNESCO) Soil Map of the World is a soil map of the entire world mapped at a 1:5,000,000 million scale that provided the first real overview of soils globally (FAO-UN - Land and Water Division (CBL), 2007). Since then, soil and soil property maps such as SoilGrids and the FAO Global Soil Organic Carbon Map (GSOCmap) fueled by global soil point datasets such as the Harmonized World Soil Database (HWSD) have improved and added to the global overview of soil information(FAO, 2020; ISRIC, 2020; Nachtergaele et al., 2009). Looking at a more regional extent other soil maps focus within their respective countries' borders such as the Digital Soil Map of the Netherlands, SOil and TERrain Database (SOTER) in Southern Africa, and the Australian Soil Resource Information System (ASRIS).

The most actively used set of soil maps within the United States is the National Soil Survey Geographic Database (SSURGO). SSURGO is maintained and updated by the USDA's Natural Resources Conservation Service (NRCS). SSURGO is a collection of digitized soil maps that are available in vector (polygon) and raster formats as well as a series of related tables containing landscape information, soil properties, usage restrictions, and soil classification information for the soil map units within the maps. Even though the SSURGO database is a rich collection of useful soil information, all SSURGO maps have soil units mapped as consociations,

complexes, and associations. Consociations are map units dominated by a single component where the remaining soils are similar and do not affect the overall interpretation of the soil (Soil Science Division Staff. 2017). Associations are two or more major soil components that occur in a regular repeating pattern within a soil map unit. Often, components of an association could have been delineated into individually mapped units, but subdivision was impractical at the time of the original mapping. Complexes, thirdly, contain different components that can occur in a regular or irregular pattern, however not on a traditionally mappable scale (Soil Science Division Staff. 2017). Within the context of SSURGO, this means that there are almost always several soils mapped within one polygon unit on the map. The combination of these soils essentially decreases the map value since it is not clear which soil type can be encountered where in the polygon without resorting to field inspection aided by the written description of the soil map unit.

The need for more detailed maps has been increasing with the advent of more complex soil modeling studies, precision agriculture techniques, and agroforestry (Bobryk et al., 2016; Jiang et al., 2015; Jin et al., 2017). Precision agriculture involves the use of subsections of fields referred to as management zones and the current resolution of SSURGO data is not adequate at this resolution (Bobryk et al., 2016). In agroforestry such as Jiang et al. (2015) SSURGO soil data is used as inputs for predictive models of site productivity for different tree species and the current resolution does not fully support this modelling. These fields would benefit from more detailed and higher resolution soil data.

One potential solution for this growing problem is disaggregation of existing maps. Generally, disaggregation refers to breaking down existing map units into smaller units (Figure 1). Disaggregation can be extremely powerful in increasing data resolution and often can be

completed with already existing data. This ability to improve the potential utility of data is a useful technique and as a result a wide variety of disaggregation techniques have been developed (Nauman et al., 2012; Odgers et al., 2014; Stoorvogel et al., 2017). Disaggregation results can range from a raster of specific soil properties to maps that assign soils to specific landscape features. This thesis is concerned with the latter.

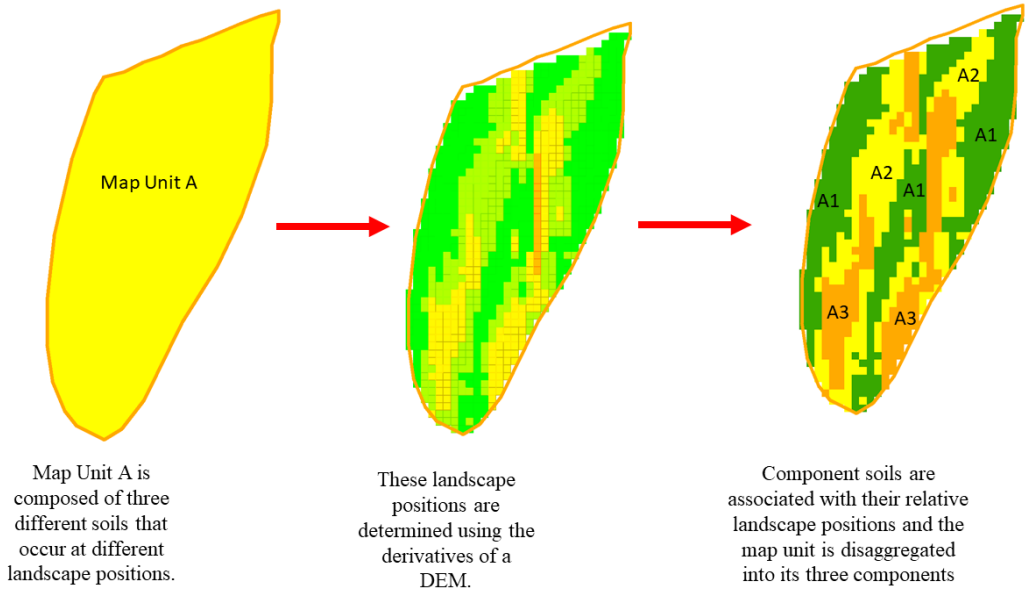


Figure 1: Simplified disaggregation procedure using landscape position to delineate components

Within soil mapping, disaggregation typically is accomplished using the existing descriptions of the relations of individual soils described within an association or complex to other known characteristics of the landscape. Different soil types within a single map unit are mostly organized by geographic positions and these geographic positions of soil in theory, can be distinguished using derivatives of a Digital Elevation Model (DEM) such as slope and curvature (e.g. Stoorvogel et al 2017). Most disaggregation methods in soil science revolve around this concept but result in two distinctly different outcomes. Polygon preserving methods such as those used by Häring et al., (2012) preserve the original map units and disaggregate within the

existing units. Conversely, Continuous raster methods such as the Disaggregation and Harmonization of Soil Map Units Through Resampled Classification Trees (DSMART)algorithm and POLARIS create a continuous prediction of soils and their properties in a raster surface (Chaney et al., 2016; Odgers et al., 2014).

While the spatial resolution of SSURGO soil map units currently does not fully meet user needs, the written descriptions attached to these units provides a higher resolution view of the soils. The main remaining unknown is how well using the descriptive information attached to these units works to disaggregate the polygons into a finer resolution and achieve better map accuracy. In terms of the original boundaries, preserving them is not inherently necessary provided they can be reconstructed to be merged into the existing digital soil map database structure. Additionally, it is important to note that many times the descriptions between different soil types can overlap because they occupy similar landscape positions. This may complicate approaches to disaggregate map units.

My objective in this thesis is to develop, implement, and test a SSURGO-tailored method to disaggregate existing soil map units based on slope steepness and other landscape position information contained within map unit descriptions across three locations in the North Central region of the United States. I will seek to answer three research questions. First, I will evaluate whether this approach is practical and efficient given the native data structure and computational limitations. Second, I will determine the impact of this process on map accuracy by comparing existing independent soil measurements in the field to the predictions in both the original SSURGO maps and the newly disaggregated maps. Lastly, I will look for regional variation of error across the different locations to determine if there are regional differences that assist or hinder the disaggregation process.

If the method is evaluated successfully, this model will be included in software package for NCRS to use to disaggregate polygons on the national level. Currently, NRCS is actively working on completing the soil map for the entirety of the United States and updating existing soil maps that are outdated or at too low resolution. A successful package resulting from this thesis ideally would allow NRCS soil scientists to further disaggregate components within map units using legacy data that are widely present and available. Additionally, the process can provide support in the creation of new soil map unit rasters by helping to identify component locations within the map units.

Background

To create and test a potential disaggregation methodology it is critical to understand both the NRCS soil surveying methodology and the state of traditional and digital soil mapping and disaggregation. Figure 2 outlines the order and structure of the background section as well as the relationships between the different types of data within traditional and digital soil mapping and disaggregation procedures.

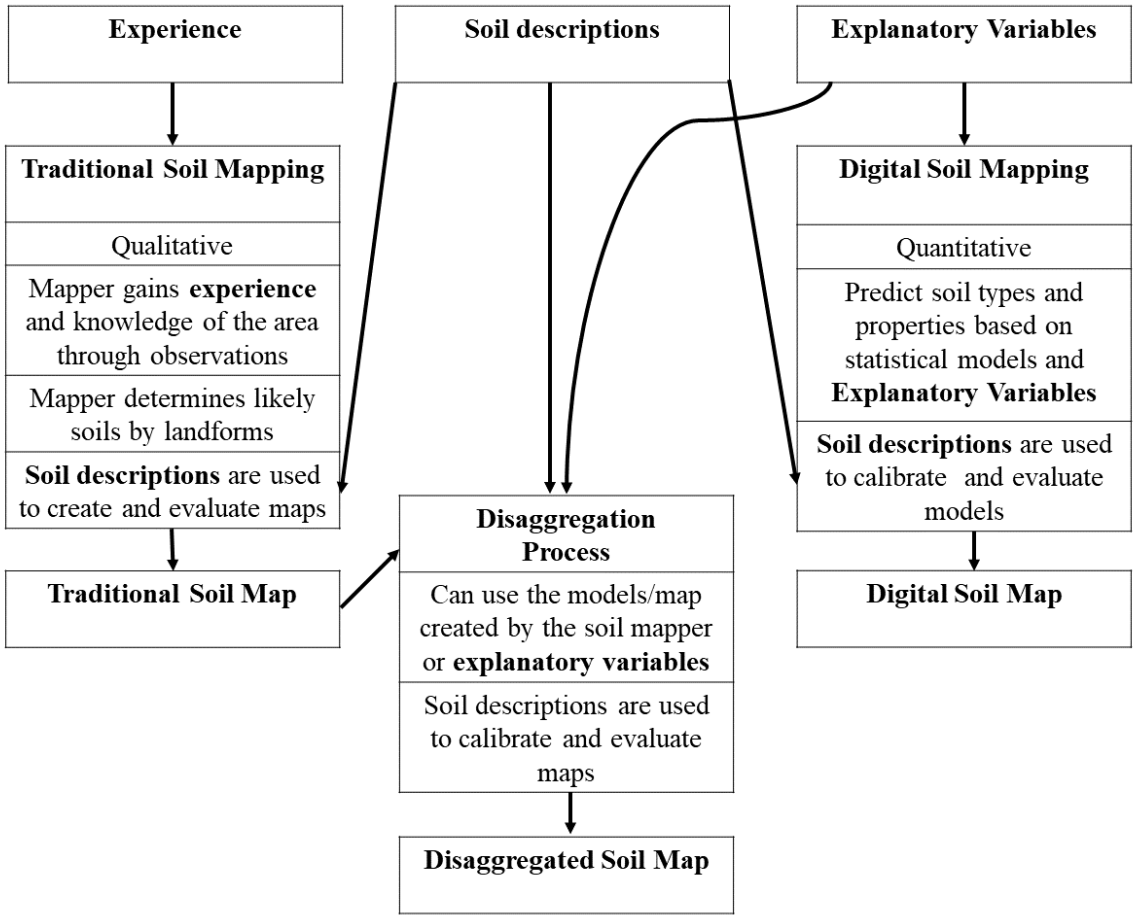


Figure 2 The relationships among traditional soil mapping, digital soil mapping, and disaggregation techniques.

Traditional Soil Mapping

Traditional soil mapping began before the advent of modern technology and had to be completed without satellite spectral imagery, advanced computer resources or predictive soil models. Therefore, much traditional mapping projects were completed by the tacit knowledge of the mapper and series of observations of different soil pedons at various locations throughout a landscape, a process that required both extensive training and long experience (Hudson, 1992). As outlined within the Soil Survey Manual (2017), traditional soils surveys were and are completed by the mapper first delineating landscape bodies into landforms. This delineation identifies areas with similar soil forming factors and similar catenas. The five soil forming factors are Climate (CL), Organisms (O), Relief (R), Parent Material (P), and Time (T) and the interaction of these 5 soil forming factors (CLORPT) leads to the formation of a particular soil (Jenny, 1941). Catenas are soil landscape conceptual models that capture how different soils occur in a predictable pattern along a transect of a landscape (Milne, 1936). This is because areas with similar soil forming factors will produce the same soil and when all are constant except the relief (R), soils can be predicted by their relative location in the landscape (Soil Survey Staff, 2017).

Next the mapper makes a prediction about the pattern of soils within this delineation based off previous knowledge of the area. The knowledge the prediction is based on is the result of thousands of hours of describing and mapping soil within the mapper's region. Provided that this knowledge and prediction is accurate, only a few strategic point observations are needed to create an adequate soil map (Hudson, 1992; Soil Survey Staff, 2017).

Third the mapper tests the delineation prediction by sampling pedons through pits, augering, and naturally occurring soil exposures. Finally, based on these observations the soil scientist either confirms the original model or rejects it and creates a new one to be tested.

If the accuracy of the soil map and its underlying hypotheses for soil distribution within landforms are found adequate, the product is certified and published. A typical final map contains delineated map units (polygons) based on the models created by soil scientists and provides support for a variety of agricultural, construction, and scientific endeavors. Note that at this completed state, still several different soil types are described within a single soil map unit (polygon), as consociate, complex, etc. Each one of these soils represents a different soil series within the map unit. A soil series is a set of soils with similar or the same profiles, parent material, and climate that have been identified and explicitly described by soil scientists. The expected location of each soils within a soil map unit polygon is provided as a textual description. Thus, a traditional map such as **Error! Reference source not found.** provides a eneral overview of the likely soils within an area, but it does not necessarily provide the visual or geospatial resolution necessary for all contemporary analyses.

Digital Soil Mapping

Moving into a modern age it became possible and necessary to add quantitative elements to soil mapping. Some early steps included purely spatial approaches that attempted to predict soils based purely on spatial relation to observations (McBratney et al., 2003). This improved to become more regionalized in the case of kriging and even implementing several covariates within the local area in the case of cokriging. Soon GIS techniques improved to allow for more geostatistical and computationally intensive techniques. This culminated in a new model

proposed by McBratney et al. (2003) that consisted of the original soil forming factors of climate (C), organisms (O), relief (R), parent material (P), and age/time (A) but added two new factors soil (S) and spatial locations (N). This new model, SCORPAN, formalized all the previous approaches into an overarching framework. Adding a soil covariate (S) allowed for soil properties to be predicted using other better known or available soil properties. Adding a spatial covariate (N) allowed for spatial coordinates and relative position to be used within the model like the previous purely spatial approaches. This new field, Digital Soil Mapping, unlike Traditional Soil Mapping, estimates a quantitative predictive model of soil types and properties that uses observation point data, but has only limited abilities to include the previously described knowledge of soil-landscape patterns, acquired by decades of fieldwork by soil experts (Figure 2).

Digital soil mapping is actively implemented within the United State and even with SSURGO data. Some digital soil mapping approaches are used to predict soil type such as Brungard et al., 2015 method in which they predict soil taxonomic classes in areas in New Mexico, Wyoming, and Utah. In this case they used machine learning methods to implement different digital soil mapping methods. Others focus on deriving individual soil properties using digital soil mapping methods such as Simbahan et al., 2006 methods to map soil organic carbon stocks in fields in Nebraska. Other methods take it even further and remap SSURGO data for the entire United States using into continuous raster surface of soil series probabilities such as POLARIS (Chaney et al., 2016).

Digital soil mapping requires extensive amounts or high-density soil observation point data to estimate and evaluate predicted soil types or properties. The legacy data-based disaggregation method proposed in this thesis contains some similarities to digital soil mapping.

However, it avoids this intensive data need. Additionally, the soil data is not being used to model soil types or properties, but instead assign more specific geographic locations to the soils described in the map units.

Disaggregation Techniques

The disaggregation work within soil mapping reflects elements of both traditional and digital soil mapping (Figure 2). Some disaggregation methods work with traditional soil maps boundaries (Häring et al., 2012). Other methods are closer to digital soil mapping and use legacy data merely as covariates (S in SCORPAN) or as calibration data for their quantitative models such as Odgers et al., (2014) and Stoorvogel et al., (2017).

One early set of methods such as those used by Bui & Moran (2001) compared three disaggregation models to discern which was the most effective to create a continuous raster surface form existing soil units in Australia. The first model used a restructuring element based on landscape relief classes calculated from the DEM and then related these to the soil position description attached to the units. The second model used a clustering method (k-means algorithm) based on vegetation coverage associated with the described soils as well as a slope map to disaggregate the soil units. The third model used a decision tree approach allowing for the use of both quantitative and categorical variables such as lithology, DEMs, relief, transport energy, and LANDSAT data. They found the best method was heavily dependent on the region and the preexisting soil information available.

In a more traditional boundary-preserving approach, Häring et al., (2012) focused on drawing new boundaries within polygons instead of creating a new continuous raster map of soil types in south east Germany. A data-driven random forest method was used to calculate probability of a soil occurring at a location with 7 terrain attributes: topographical wetness index,

relative height floodplain index, a modified floodplain index, a mass balance index, slope gradients, mid-slope positions. A soil-designation threshold of a probability value above 0.7 was used. Any location where no soil reached a probability of presence above 0.7 was classified as "indifferent." Haring et al found that slope, along with flood plain index values and relative height were the strongest predictors in the model (2017).

Other methods push even further into the data driven digital soil mapping and SCORPAN methodologies. SCORPAN's use in the disaggregation of soil map units is best seen with the Disaggregation and Harmonization of Soil Map Units through Resampled Classification Trees (DSMART) (Odgers et al. 2014). In this approach continuous raster surfaces were created using a series of SCORPAN covariates, legacy information from the map unit descriptions, and decision trees. Subsequent approaches such as Vincent et al., 2018 expanded on this approach in Northwestern France. As inputs for the algorithm, they used landscape position and "expert knowledge" from soil surveyors as well as environmental covariates such as geology, land use, and terrain attributes. Within the decision trees parent material and original map polygons were used 100% of the time and other covariates were used less. Overall accuracy varied from ~40% to ~ 70% showing some success for this approach. Even though original soil map polygons were used as a covariate this did not prevent those soils from existing outside the boundaries.

Other digital soil mapping approaches to disaggregation include Stoorvogel et al. (2017) S-world approach to disaggregate polygon-based soil maps using expert knowledge and simple statistics (2017). However, in this case the goal was to generate soil property maps not disaggregate soil map units into their individual soil types. In this approach Stoorvogel et al. accomplished the disaggregation of map units by starting from the bottom of topographic

sequences of soils and assigning soils sequentially by their relative topographic positions pulled from the soil legacy data.

Some approaches such as Nauman and Thompson (2014) and Chaney et al. (2016) use SSURGO data within their disaggregation. Nauman and Thompson wanted to disaggregate adjacent surveys into a single universal map without the original survey boundaries. Their disaggregation procedure used a supervised classification procedure based on the soil descriptions and hillslope, landform, slope, elevation, aspect, and catchment area rasters as inputs. All the descriptions were obtained from SSURGO tables. Then, based on the SSURGO soil descriptions, two soil properties (rock content and soil organic carbon percentage) were assigned a probability value and a continuous soil properties raster surface was calculated. This procedure was very time intensive, required extensive code and failed to produce accurate maps at a local scale (Nauman & Thompson, 2014). It also bears reminding that their end goal was a soil property map and not disaggregated SSURGO units.

Chaney et al., 2016 in their creation of POLARIS used the DSMART algorithm to disaggregate SSURGO data for the contiguous United States (CONUS). Their approach started with a suite of environmental covariates such as elevation, slope, curvature, parent materials, land cover, and topographic indices as well as legacy SSURGO data. Using these as inputs the DSMART algorithm randomly samples at least 100 soil observations per map unit and calibrates decision trees that predict the most likely components out of the thousands of possible components. The components are assigned a rank by likelihood for the first 50 components. Overall, at rank 1 the accuracy was around 17% for each soil series and reached 68% percent when all 50 ranks were considered.

Overall, most of these methods employ the use of the existing legacy data to calibrate or provide covariates for their models. In most cases it was to either supplement or reduce the pedon data needed within each study. Additionally, much of the success was either regionally or scale dependent. However, none of these studies have looked to update existing map units within SSURGO while still preserving the current format and structure.

Study Sites

Two locations in the north central region of the United States were selected to create and test the proposed disaggregation methodology. These are the entirety of Mason County Michigan and the Mille Lacs Region of Minnesota. These two sites cover different spatial extents, varying climates, soil types, and geologic histories.

Table 1 Overview of climate information of the two study sites. (Derived from USDA Annual Precipitation and Temperatures for the Conterminous United States)

Study Site	Area (km^2)	Average Maximum Temp (C °)	Average Minimum Temp (C °)	Average Precipitation (mm)
Mason County	1320.1	12.5	2.5	863.6
Mille Lacs	5078.4	11.1	-0.8	736.6

Mason County, Michigan

The first site, Mason County, MI, is on the eastern side of lake Michigan and covers over 1320 kilometers. The county has a marginally wetter climate on average than the Mille Lacs site and is generally cooler overall. The area is dominated by prior glacial activity and current lake Michigan activity resulting in large deposits of sand and glacial moraine till as well as active dune land especially near the lake shore. Looking broadly this area is covered by mainly four

different soil orders Entisols, Spodosols, Alfisols, and Histosols. Entisols are generally described as weakly developed, are often from recent depositions of largely unaltered parent material, and have not accumulated substantial amounts of organic matter. The Entisols present are mostly comprised of the duneland near the lake shore that have steep slopes, with fast deposition and erosion rates inhibiting major soil development. Additionally, associated with the coarse sandy material of the region is the spodosols which are soils where organic matter and aluminum are weathered into the subsoil. Alfisols are soils that experience leaching and weathering, causing the clay within the surface layers to move down into the subsoil. This allows soil moisture and other plant benefiting nutrients to be present within the subsoil and support more complex and deeper root networks. As a result, they generally occur with forest or other relatively heavy vegetative cover. Alfisols cover most of the county, dominantly in areas that are at relatively higher elevations than the lake. Histosols are extremely organic rich soils and are saturated almost throughout the year and throughout the profile. They generally occur in peatbog or swamp type settings where water content is high and organic plant remains deposit faster than they decay. Histosols mostly occur within the county in areas that are at lower elevations and within depressions.

Looking more specifically the area is dominated by very sandy soil series resulting from the lake and prior glacial activity such as the Cover, Grattan, Kingsville Mucky, and Spinks-Coloma sands (Soil Survey Staff, 2020). These series are present in over 181.3 km^2 or just under 14% of the entire county. The series differ mostly in their drainage ability. The Grattan Sands are excessively drained while the Covert series have more moderate drainage and the Kingsville is very poorly drained. The Spinks Coloma series is associated with dunes and drifting sands and are distinguished from one another by the absence or presence of clay accumulation in the

subsoil. Other common soil series within the county includes the Wixom-Capac Complex which formed from sandy and loamy till deposits and covers about 49.2 km^2(Soil Survey Staff, 2020). They are distinguished from another by the presence of coarse sandy material in the upper horizon, the Wixom, or the lack of this sandy material, the Capac otherwise they occupy similar landscape positions. Occupying similar landscape positions and being distinguishable only by horizon differences may provide issues in terms of this approach because it relies on landscape position.

Overall, Mason county consists of about 75.6 % consociations, 21.6 % complexes, and the other 3 % is undifferentiated groups. Since Mason county has many different townships there are some artificial boundaries and previous survey borders present within the soil map (Figure 3). However, the entire county is mapped at order 2: the second highest order of soil mapping.

Mille Lacs Region, Minnesota

The second study site, the Mille Lacs Region, is significantly larger covering about 5078.4 $km^2$and spanning over 9 counties in Minnesota. The Mille Lacs region is colder and drier on average. Like Mason County, the region is dominated by prior glacial activity which has impacted the topology and geology for the area. However, in the Mille Lacs Region the glacial activity was a series of alternating glacial retreats and advances resulting in glacial till as a prominent parent material within the soils in the area. The major soil orders within the areas are Alfisols, Entisols, and Histosols, and are identical to those present in Mason County. This reflects the similar landscape and vegetative features such as coastal beach and dune activity, swamps and bogs, and forest cover. Looking more specifically the major soils series within the area consist of the Mora-Ronneby complex that covers 400.5 km^2, the Milaca 164.6 km^2 and the

Milaca-Millward complex 109.8 km^2, the Seelyville 122.7 km^2, and the Rifle 99.1 km^2 (Soil Survey Staff, 2020). The largest soil series within the map units, the Mora-Ronneby Complex, occurs on glacial drumlins and moraines at 1 to 7 % slopes. Both the Mora and Ronnebysoil series have a densic contact at around 100-150 cm and are mostly distinguished by landscape position with the Mora series occupying shallower slopes or lower elevations and the Ronneby series occurring in flats or depressions. The Milaca and Millward series are like the previous soils forming on glacial drumlines and moraines and having a rocky contact at around 100-150 cm from the surface. However, these soils occur at steeper slopes than the Mora or Ronneby. The Millward series within the complex units is distinguished from Milaca by a sandy layer between the surface and densic rock contact layers. The other soil series most common within the area are the Rifle and Seelyeville series and they reflect the swampy fraction of the Mille Lacs region. The Rifle series is characterized by poorly drained peat with more than 130 cm of organic material in depressional areas. The Seelyeville is very similar in positions and drainage characteristics however, it is not characterized as peat and is instead classified as muck. Within soil science muck is generally classified as sapric, less fibrous , while peat is classified as fibric, mostly fibrous (Soil Survey Staff, 2014). Overall, the Mille Lacs region consists of about 65.2 % consociations, 30.6 % complexes, 0.2 % associations, and the other 4 % is undifferentiated groups. Like Mason County, the Mille Lacs region also contains artificial discontinuities and borders within the original map especially along county lines. However, unlike Mason County only 38.4 % is classified as order 2 with rest being older and unclassified.

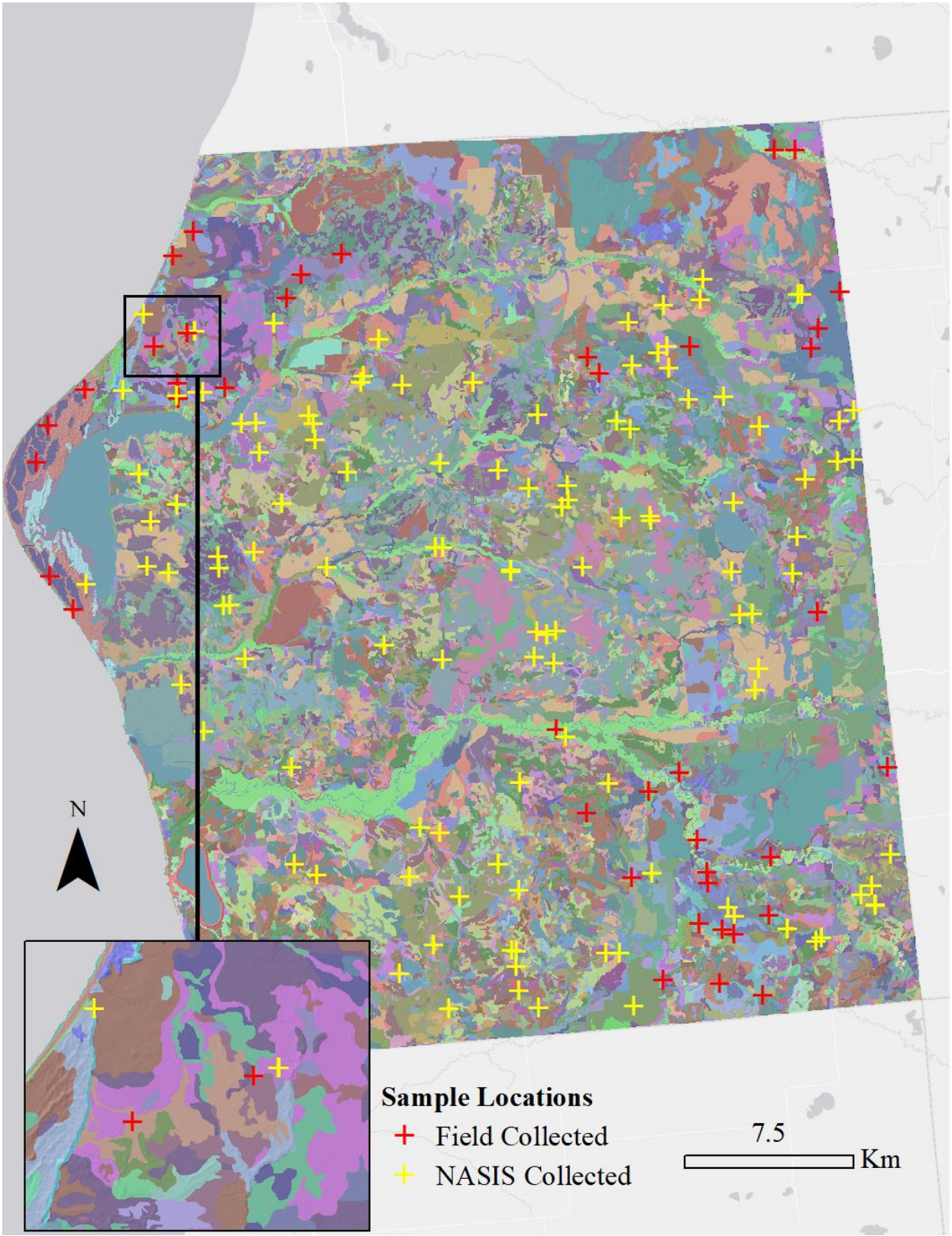


Figure 3 The gSSURGO Soil Map for the Mason County, Michigan Study Site and the sample locations for the evaluation data.

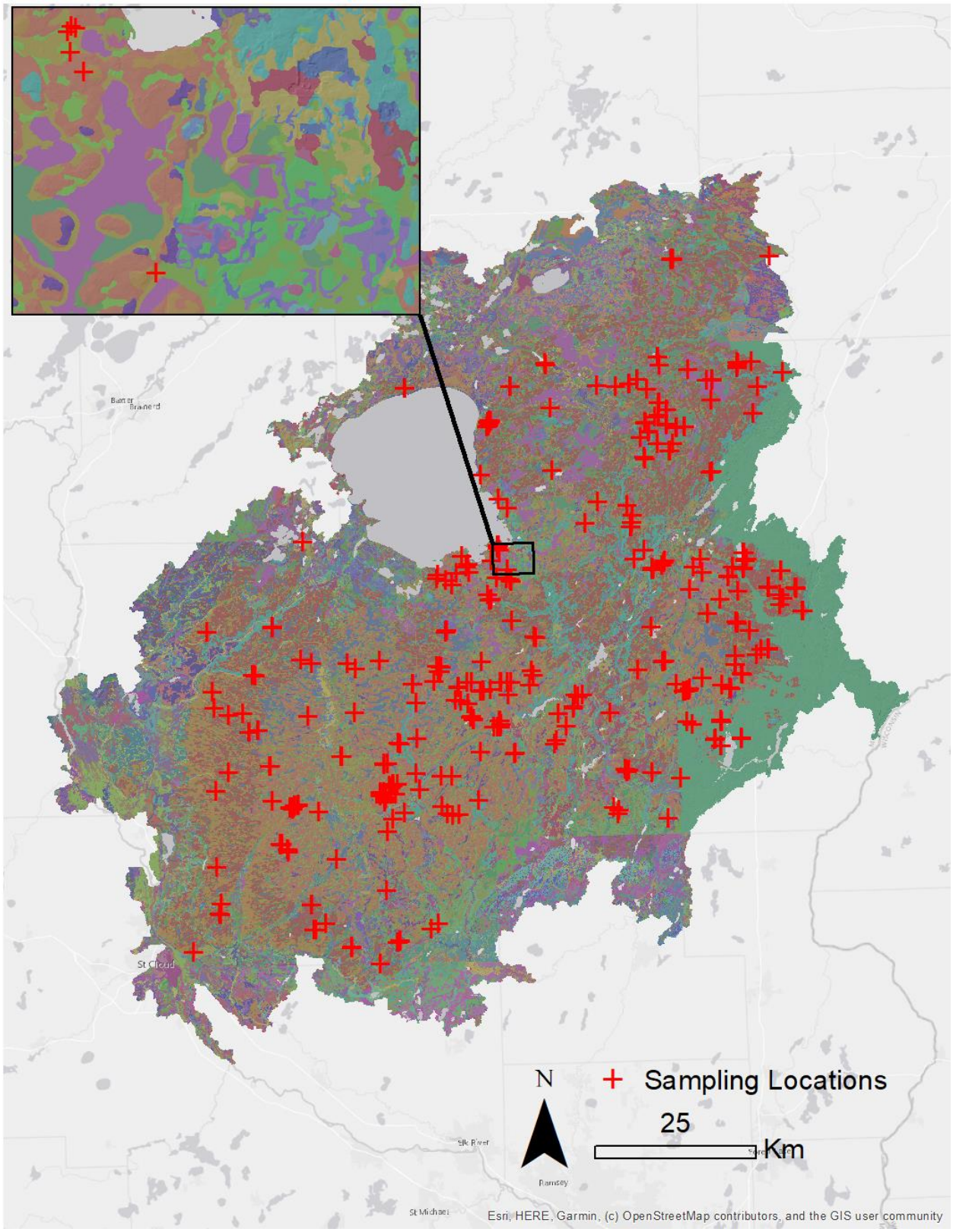


Figure 4 The gSSURGO Soil Map for the Mille Lacs, Minnesota Study Site and the sample locations for the evaluation data.

Methods

The method within this thesis requires a variety of data inputs that need to be integrated at different steps within the process. The data used and its order within the method is outlined in

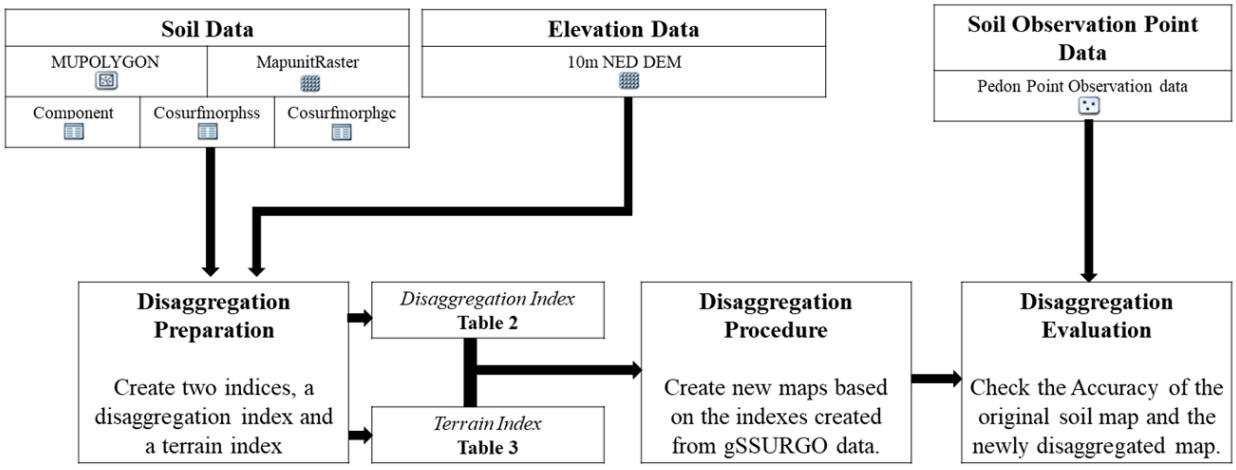


Figure 5 Overview of the methods and evaluation within this thesis, showing the input soil and elevation data, the preparation steps leading into the procedure, and the evaluation of the original and produced maps.

Soil Data

The starting soil data was sourced from the Gridded Soil Survey Geographic Database (gSSURGO) through the USDA/NRCS: Geospatial Data Gateway as a file geodatabase for each state. The soil geodatabase contains extensive soil information, but only the tabular component, slope, curvature, and landform information, a map unit polygon feature class, and the soil map unit raster at 10-meter resolution was extracted (Figure 5). The two feature classes, the map unit polygon and soil map unit raster, were clipped to the respective study area. The tables were joined by their related fields and condensed to individually mappable components that are distinguishable by slope or curvature, and that consist of one or several soil series. This newly generated table, the Disaggregation index (Table 2), provides information about which components of soil map units can be disaggregated as well as the landscape positions under

which these components are found. In some cases, two components cannot be separated from each other based on the landscape position. Where this happens those two components are merged in the disaggregation index and will form on soil map unit. The index contains all the necessary input slope and curvature information to set the fuzzy functions targets, retains the component keys and area percent information, and all the original map unit keys. An example fraction of a Disaggregation Index is shown in Table 2.

Table 2 A fraction of the Disaggregation Index for Mason County with the slope information from the descriptions such as slope low, representative, and high values. The table also contains curvature information such as planar curvature (Shape Across) and profile curvature (Shape Down). Finally, the table contains the first and last component keys and their respective area component percentages within an existing SSURGO map unit. Illustrating this with the third row in the index for example, the unit-to-be-created is currently part of existing SSURGO map unit key 1382763. It still consists of two components that cannot be disaggregated based on SSURGO legacy descriptions (namely, 18311190 and 18311193). The unit-to-be-created is one of three creatable new units within the original map unit 1382763 and differs from the other two (in the first two rows of the index) in terms of its slope range. Specifically, the slope range extends to 2% slopes: the highest of the three creatable units within that map unit. The two components in the creatable new map unit together make up 10 percent of the existing map unit.

Map Unit Key	Slope Low Value	Slope Representative value	Slope High Values	Shape Across	Shape down	First Component Key	Last Component Key	Component Percentage
…	…	…	…	…	…	…	…	…
1382763	0	0	1	Linear	Linear	18311191	18311191	5
1382763	0	1	1	Linear	Linear	18311192	18311192	85
1382763	0	1	2	Linear	Linear	18311190	18311193	10
1382764	0	0	1	Concave	Concave	18311146	18311146	1
1382764	1	2	3	Convex	Convex	18311147	18311147	9
1382764	1	2	3	Linear	Convex	18311145	18311145	90
1382765	0	0	1	Concave	Concave	18311075	18311075	1
1382765	1	2	3	Convex	Convex	18311077	18311077	7
1382765	2	7	12	Linear	Convex	18311076	18311076	2
1382765	3	5	7	Linear	Concave	18311078	18311078	3
1382765	3	5	7	Linear	Convex	18311074	18311074	85
1382765	3	5	7	Linear	Linear	18311073	18311073	2
…	…	…	…	…	…	…	…	…

Elevation Data

The DEM data was sourced from the National Elevation Dataset (NED) through the Geospatial Data Gateway and is at 10 m resolution (U.S. Geological Survey, 2015). This resolution was chosen because it is available across the United States, including the study sites, and it matches the resolution of the Soil map raster within the gSSURGO geodatabase. Next, DEM derivatives were calculated consisting of the original slope, planar curvature, and profile curvature (Figure 6). To obtain a higher level of generalization to complement the 10 m resolution, the DEM was aggregated by a factor of 5 to a 50-m resolution. Slope, profile curvature, and planar curvature were calculated using standard ArcGIS tools for both the 10-m and 50-m DEM. After these six terrain attributes were calculated, a focal mean filter was passed over each of the results. At this point, twelve rasters with terrain attributes were available for further analysis (Figure 8).

During the disaggregation procedure, these twelve terrain attributes were used to have some control over the patterning of the disaggregated map. For example, weighting the aggregated version of the terrain information higher will cause the resulting map to coarsen because each 5 by 5 block of cells has been converted to a single value. Conversely in the focal mean the terrain information is averaged for each cell in relation to the 8 cells surrounding the cell. Unlike the aggregated version each cell will keep an individual value however, it has been averaged compared to the surrounding cells and will smooth the resulting map. The information from the 12 attributes was saved in one table, the terrain index, with the number of rows equal to the number of cells in the study area, and the number of columns equal to the number of DEM derivative variants (Table 3).

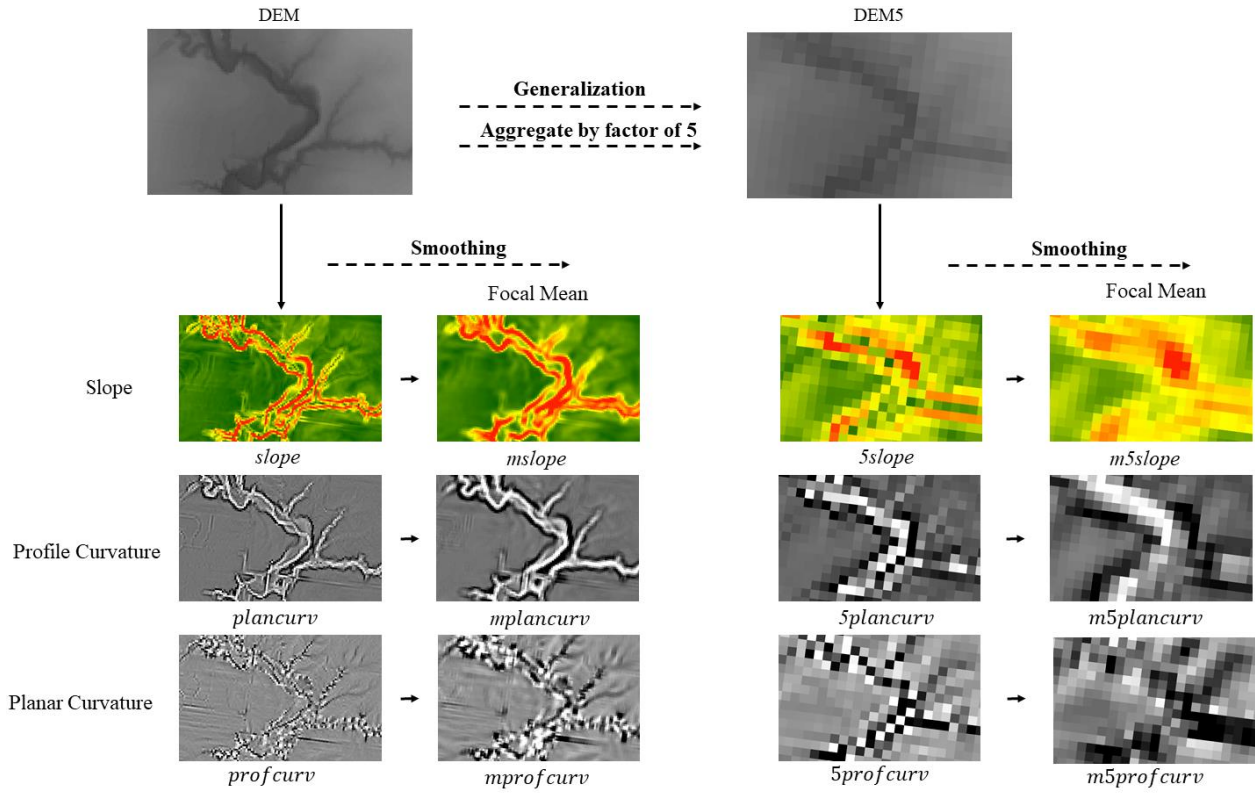


Figure 6 The 12 DEM derivative variants following from two DEMs with different resolutions, with and without smoothing using the Focal Mean operation. These serve as inputs during the disaggregation procedure and allow control over the resolution and smoothness of the disaggregated soil map

Table 3 A fraction of the Terrain Index Table for Mason County. It contains values for all of the 12 DEM variants for every cell in a study area as well as the map unit key referencing the originally mapped soil map unit. Note that aggregated values calculated over larger resolutions are repeated several times at the smaller resolution of the terrain index.

Map Unit Key	Slope	Focal Mean Slope	Aggregated Slope	Focal Mean Aggregated Slope	Planar Curvature	Focal Mean Planar Curvature	Aggregated Planar Curvature	Focal Mean Aggregated Planar Curvature	Profile Curvature	Focal Mean Profile Curvature	Aggregated Profile Curvature	Focal Mean Aggregated Profile Curvature
2605773	0.552	0.539	83.562	57.221	-0.020	-0.010	-17.938	-2.483	0.015	-0.011	-18.179	-57.798
2605773	0.537	0.513	83.562	57.221	-0.030	-0.017	-17.938	-2.483	0.024	-0.010	-18.179	-57.798
2605773	0.462	0.458	83.562	57.221	-0.020	0.014	-17.938	-2.483	-0.014	-0.011	-18.179	-57.798
2605773	0.344	0.440	83.562	57.221	0.074	0.025	-17.938	-2.483	0.009	-0.008	-18.179	-57.798
2605773	0.432	0.395	83.562	57.221	0.050	0.034	-17.938	-2.483	-0.023	0.005	-18.179	-57.798
2605773	0.299	0.363	11.639	38.886	0.022	0.019	0.514	-3.511	0.026	-0.002	-0.215	-19.197
2605773	0.279	0.363	11.639	38.886	0.030	0.021	0.514	-3.511	-0.047	-0.010	-0.215	-19.197
2605773	0.416	0.459	11.639	38.886	0.029	0.036	0.514	-3.511	-0.003	-0.032	-0.215	-19.197
2605773	0.604	0.654	11.639	38.886	0.032	0.026	0.514	-3.511	-0.059	-0.035	-0.215	-19.197
2605773	0.915	0.797	11.639	38.886	0.034	0.023	0.514	-3.511	-0.063	-0.016	-0.215	-19.197
2605773	0.890	0.798	21.521	27.036	-0.022	0.013	0.183	0.553	0.074	0.015	-0.501	0.028
2605773	0.616	0.679	21.521	27.036	-0.008	0.005	0.183	0.553	0.041	0.025	-0.501	0.028
2605773	0.552	0.697	21.521	27.036	0.015	-0.001	0.183	0.553	-0.036	-0.034	-0.501	0.028
2605773	0.955	1.005	21.521	27.036	-0.019	-0.008	0.183	0.553	-0.118	-0.075	-0.501	0.028

Disaggregation Procedure

It is now possible to disaggregate the components similar to what is seen in Figure 1. The soil map units were disaggregated based on their recorded landscape positions using the Disaggregation Index created from the information in the gSSURGO database, the 12 DEM terrain derivatives in the Terrain Index, and the gSSURGO 10-meter resolution soil map unit raster. However, as visible in Table 2, soil map components *cannot be directly assigned to areas of a certain slope or curvature values because they are described in a range around their recorded values* or in the case of curvature, they become more likely as the convexity or

concavity increases or decreases. Additionally, slope ranges and curvatures for multiple new soil map units often partially overlap making it impossible to directly assign new soil map units to a location. To account for this difficulty, gaussian fuzzy functions are used to assign scores to each component based on *how strongly* (instead of *whether*) the location fits the descriptions of each disaggregated soil map unit (Figure 7). Using a fuzzy function allows for a smooth transition away from the most probable values of slope and curvature instead of a binary yes or no assessment. The partial probability, μ, of soil in a location is thus predicted using a gaussian fuzzy function such as Equation *1*:

$$\mu(x)_{s,i} = e^{-f1*(x_{s,i}-f2_{s,i})^2} \qquad 1$$

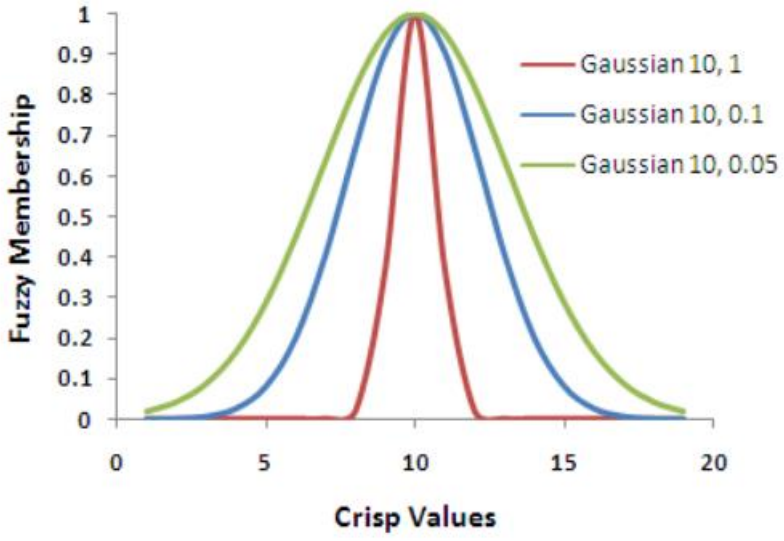


Figure 7 Example of three gaussian functions with a center value (*f2*) of 10.

The gaussian fuzzy functions consist of three main elements, the input x or in this case the slope/curvature value, the spread *f1*, and the midpoint *f2* or the value target. For instance, in the case of slope, x_s is the value of slope at a given cell, s, within the map. *f2* is the target value of

slope for the component being assigned. As x approaches the f2 value the output approaches 1. The *f1* values determines how quickly the output values moves away or towards the value of 1. The Gaussian fuzzy functions midpoints (*f2*) were set using the expected component slope ranges and curvature values pulled from the Disaggregation Index.

The terrain information was inserted as x into the function and was pulled from the Terrain Index. Finally, a constant (0.1) was set as the spread (*f1*) and the relative likelihood that a soil is found within a single raster cell for a particular terrain attribute is determined. The likelihoods are not probabilities but scores that are relative to the likelihoods of the same original map unit soils within that unit.

The 12 fuzzy function outcomes – one for each version of each terrain derivative - are combined into an overall soil component score that expresses the partial relative likelihood of each possible soil component *i* for a raster cell *s* (Equation *2*):

$$Partial\ Soil\ Component\ Score_{s,i} = \qquad 2$$

$$\alpha * \mu(slope)_{s,i} + \beta * \mu(mslope)_{s,i} + \gamma * \mu(5slope)_{s,i} + \delta * \mu(m5slope)_{s,i} + \alpha * \mu(plancurv)_{s,i} + \beta * \mu(mplancurv)_{s,i} + \gamma * \mu(5plancurv)_{s,i} + \delta * \mu(m5plancurv)_{s,i} + \alpha * \mu(profcurv)_{s,i} + \beta * \mu(mprofcurv)_{s,i} + \gamma * \mu(5profcurv)_{s,i} + \delta * \mu(m5profcurv)_{s,i}$$

$$\alpha + \beta + \gamma + \delta = 1 \qquad 3$$

Equation 2 is comprised of two elements: the gaussian fuzzy function results of each of the 12 DEM derivatives in location *s* for component *i* and respective weight coefficients α, β, γ, or δ that allow for control over the resulting maps shape, coarsening or smoothing. For example, increasing the value of γ would increase the weight of the aggregated DEMs therefore increasing

their impact on the resulting score. As a result, the map would coarsen because the aggregated cells would have more influence over the result.

The four coefficients, α, β, γ, or δ, add up to 1 and since the individual fuzzy scores are values from 0 to 1, the resulting score from this equation will also be between 0 to 1. These four coefficients are currently held stable for all *s* and *i* but may be varied by mappers to better reflect their understanding of a region's soils. For instance, when a component *locally* does not have a clear relation to the landscape a *locally* coarser map with high γ, or δ may provide a better estimate.

At this point, it is possible to assign the soil components with the highest score to each location (grid cell) in a map. However, this would not ensure that each component occupies the SSURGO-prescribed percentage of each original soil map unit (Table 2). [1]

To gain control over the relative area percentages predicted for each component *i*, a calibration parameter ε_i is introduced. The parameter is used to make components less or more likely relative to other components, regardless of location *s* (Equations 4 and 5):

$$Final\ Soil\ Component\ Score_{s,i} = Partial\ Soil\ Component\ Score_{s,i} + \varepsilon_i \qquad 4$$

$$Mapped\ Soil = Max\ (Soil\ Component\ Score_{s,i}) \qquad 5$$

Using a first guess for $\varepsilon_i = .5$, the final component score is calculated for each possible component *i* in each location *s* (raster cell) and the component with the highest score is assigned.

[1] SSURGO soil map unit descriptions proscribe the relative proportion of the unit that is occupied by each component, as in Table 2"

Once every cell within a map unit has been assigned a component, the percent cover of each component is calculated, and the value is compared to the required area percent described in SSURGO and recorded in the disaggregation index. E_i is then iteratively increased if a component was predicted over less area than required and iteratively decreased if a component was predicted over more than required. This process is repeated until the area percent deviation is within a user specified range. I chose +-2% of the original soil map unit component percentage for this range because it provided a practical and reachable target and decreasing it any further did not affect accuracy assessment results. However, in some cases it was not a reachable target and the model would stop trying to reach it after the 50 iterations I allowed and assign the best result it reached.

Once acceptable assigned areas have been reached for all newly disaggregated soil map units in each original polygon, the newly disaggregated map units are assigned their component keys and are saved as new map units. This process is repeated for each map unit and the results are merged back into one raster soil map. The outcome of the procedure is therefore a new soil map raster that has disaggregated soil map units in SSURGO format and that best matches the area percentages described in the original SSURGO soil map unit descriptions.

To evaluate each disaggregated map, independent soil point pedon data were collected for each site. For Mason County a total of 167 points were described, consisting of recent randomly assigned points within complexes (41) and a set of older unused soil descriptions provided by NRCS soil scientists from the region (126) (Figure 3). The point description data set for Mille Lacs consists of 400 random points collected for various recent digital soil mapping projects and other regional mapping projects (**Error! Reference source not found.**). Each of

hese datasets are effectively random in relation to this study and provide adequate coverage of each area.

Accuracy assessment

To determine accuracies, each sampled point was checked to see if the observed soil series matched any of the predicted soil series in the new soil map unit and the original SSURGO map. Accuracy was expressed as the mean error of all *n* observations in Mille Lacs and Mason County (Equation 6):

$$Accuracy = 1 - \frac{1}{n}\sum_{i=1}^{n}(fraction_{obs,i} - fraction_{pred,i}) \quad 6$$

where $fraction_{obs}$ is the areal fraction assigned to the observed soil series, and $fraction_{pred}$ is the areal fraction of the same soil series that was predicted. In all cases, $fraction_{obs,i}$ was 1, i.e. field observations always resulted in only clear soil series, not a mix of soil series. Therefore, if all $fraction_{pred,i}$ are zero (i.e. if the observed soil series is never among the predicted soil series in any disaggregated soil map unit), then accuracy is 0.

The tables with the partial scores for both the SSURGO and the disaggregated map are provided in the appendix as **Error! Reference source not found.** and **Error! Reference source t found.** for Mason County. To better compare to other methods such as POLARIS that use a 30-meter resolution DEM the accuracy assessment was also completed using a 3 by 3 window around the sample points. In that case, the score was calculated as in equation 6 for each of the 9 cells in the 3X3 window and the highest score was kept. This essentially allowed 9 attempts for each sample point.

To help assess accuracy, assessment confusion matrices were generated at the soil order level. Confusion matrices provide a variety of insights into classification success such as user accuracy and producer's accuracy. User's accuracy reflects the number of correctly predicted samples divided by the number of samples within the row and provides a measure of reliability of the observations (Janssen & van der Wel, 1994). Producer's accuracy on the other hand is the number of correctly predicted samples divided by the total number of samples in the column. Producer's accuracy provides a measure of the percentage of a class that was correctly classified (Janssen & van der Wel, 1994). To generate confusion matrices within this project the soil order of the sample point (reference) was determined using the SSURGO tabular data and then was compared to the soil order of the component of the disaggregated map (observation). After matches were determined, user's and producer's accuracies were generated using the correctly assigned divided by the row or column total respectively.

Results

Practicality of The Approach

Because the model is designed to be a practical and efficient the procedure was run on standard office computers and the specifications are provided below (Table 4). The size of the study area greatly affected runtimes (Table 5). The runtimes for Mason County were relatively quick and ranged from just over 1 minute to create the Disaggregation Index to about 2 hours 25 minutes to create the disaggregated map. In contrast the runtimes for the largest study site, the Mille Lacs, region ranged from a little over 4 minutes for the Disaggregation Index creation to over 45 hours to disaggregate the map. This size of the files also followed this trend with the sizes for Mason County being relatively small ranging from a just over 50 kilobytes to around 6.7 gigabytes. For the larger site, the Mille Lacs region, these values greatly increased especially for the disaggregated map to over 49 gigabytes. Additionally, for the Mille Lacs region the files became too large to be run at once and needed to be split into four different parts.

Table 4 Computer specifications for the office computers used for the disaggregation procedure.

Computer Specifications	
Model	Dell Precision T3600
Processor	Intel® Xeon® CPU E5-1620 0 @ 3.60 GHz
Installed Ram	16.0 GB
System Type	64-bit operating system

Table 5 The script runtimes for each study site and their areas.

Study Site	Area (Km^2)	Disaggregation Index Creation	Terrain Index Creation	Disaggregated Map Creation
Mason County, MI	1320.1	1 m 12 s	43 m 42 s	2h 25m 51s
Mille Lacs, MN	5078.4	4 m 21 s	4 h 54 m 18 s	45h 38m 08s

Table 6 The file size of each output/input during the disaggregation procedure.

Study Site	Area (Km^2)	Disaggregation Index Size	Terrain Index Size	Disaggregated Map Size
Mason County, MI	1320.1	54.3 KB	2.31 GB	6.71 GB
Mille Lacs, MN	5078.4	234 KB	16.7GB	49.1GB

Disaggregated Maps

Disaggregated SSURGO format soil maps were created for each study area using the model and inputs described previously. Because of the size of the other study areas the figures reflect the insets established before within each map and the full maps are provided in the Appendix. Looking closer within each original map unit new disaggregated components are visible (Figure 8, & Figure 9). This is further shown by the increase in the number of smaller polygons within each map in the histograms (Figure 10). The number of polygons increased dramatically in the first several bins for the study sites.

Other additional metrics were calculated such as the difference between minimum and maximum epsilons and the deviations between the disaggregated and SSURGO-prescribed are

percentages (Figure 11 -Figure 14). The epsilon maps reflect the difference between the minimum and maximum epsilons values of the components within each original polygon. The area percentage maps reflect the difference between the newly disaggregated component area percentages and the component percentages originally listed in SSURGO.

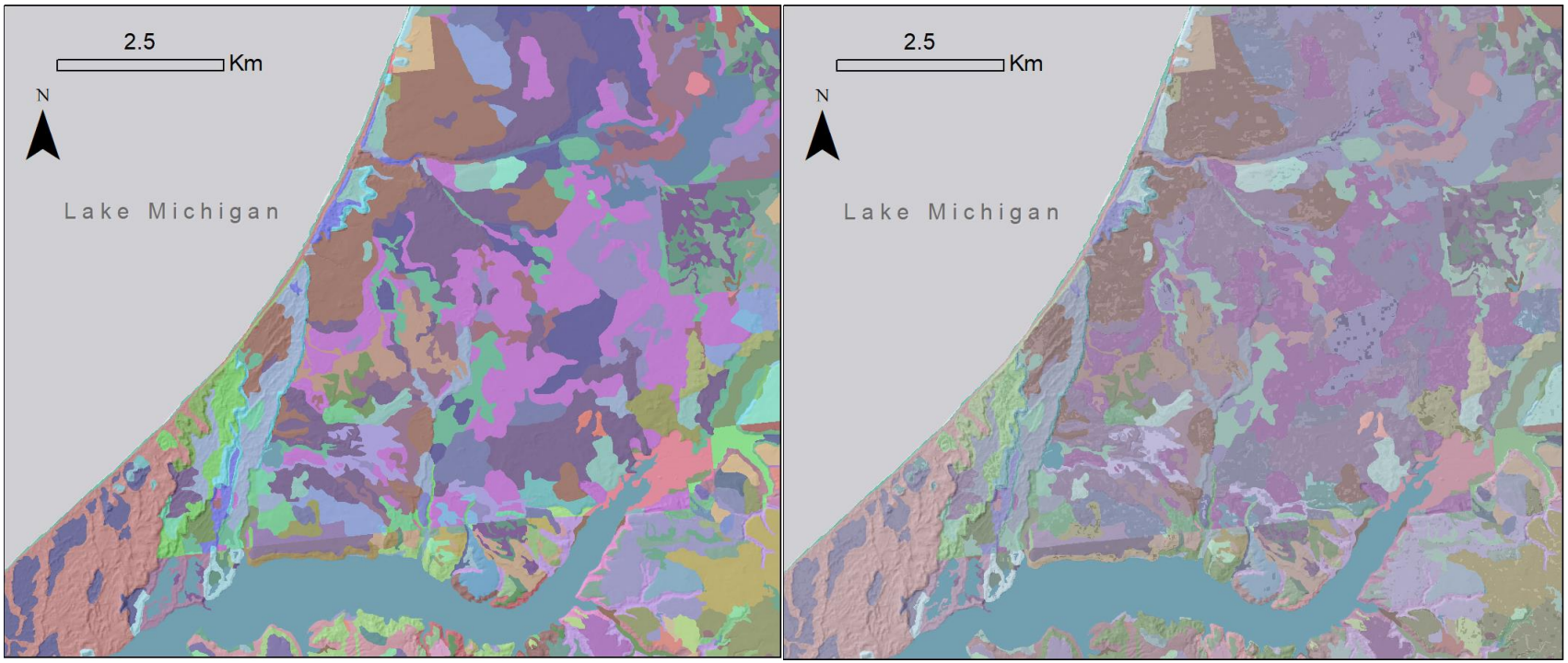


Figure 8 Before disaggregation and after disaggregation versions of the Mason County Inset Soil Map.

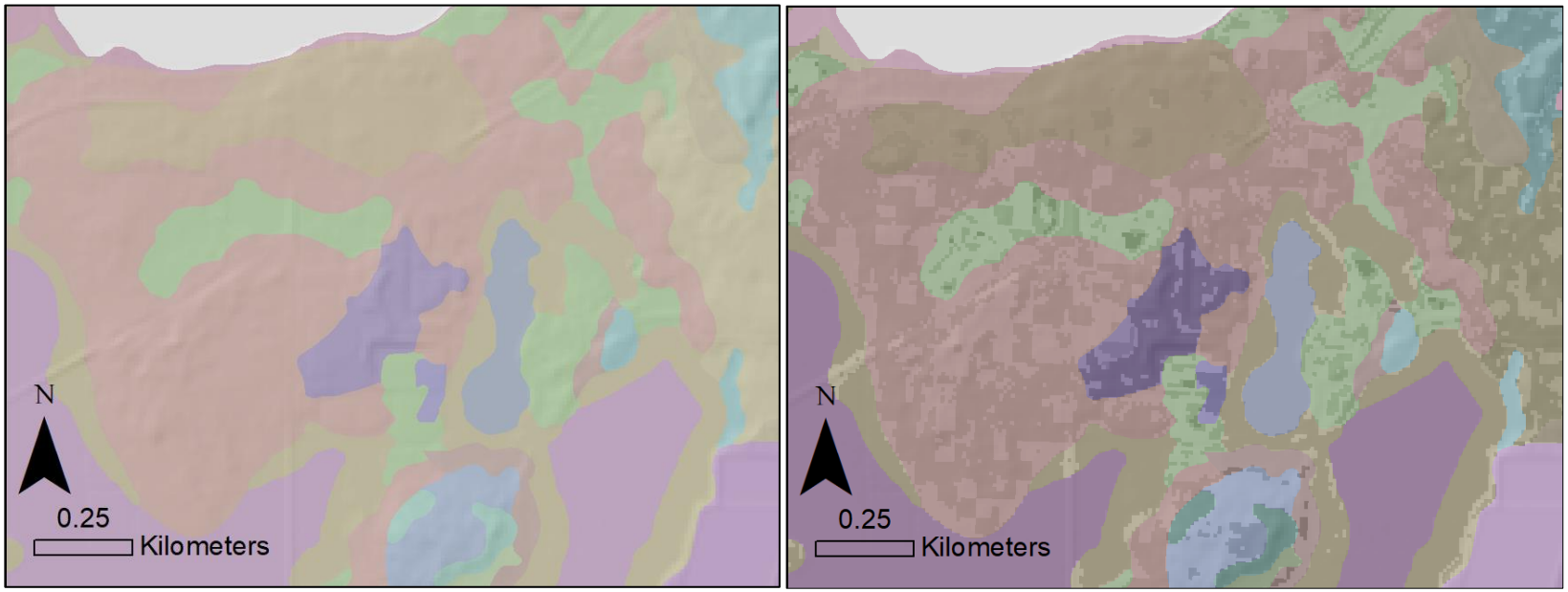


Figure 9 Before disaggregation and after disaggregation versions of the Mille Lacs region Inset Soil Map.

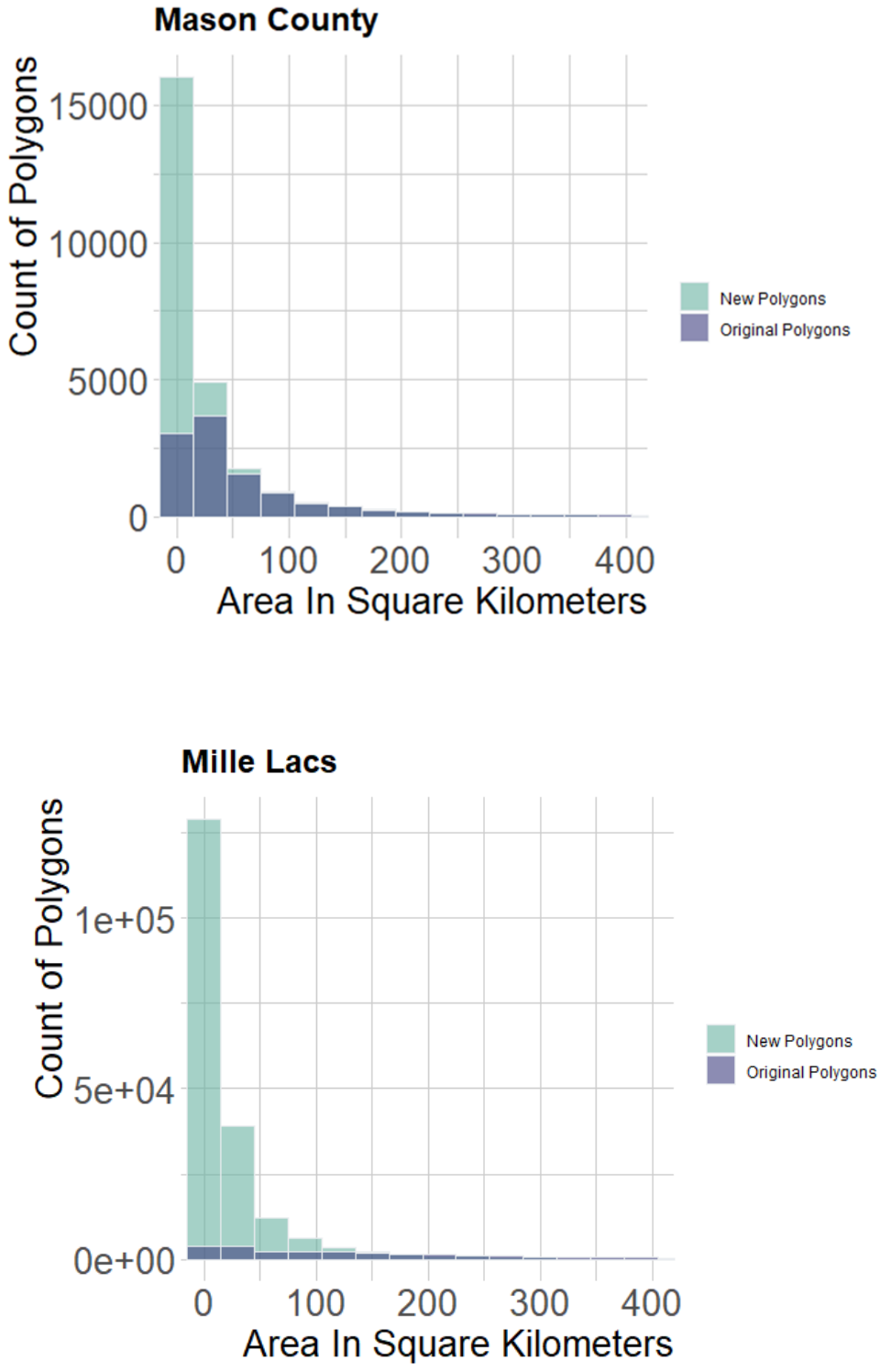


Figure 10 Histograms demonstrating the count of polygons by area. After disaggregation the number of smaller polygons increased.

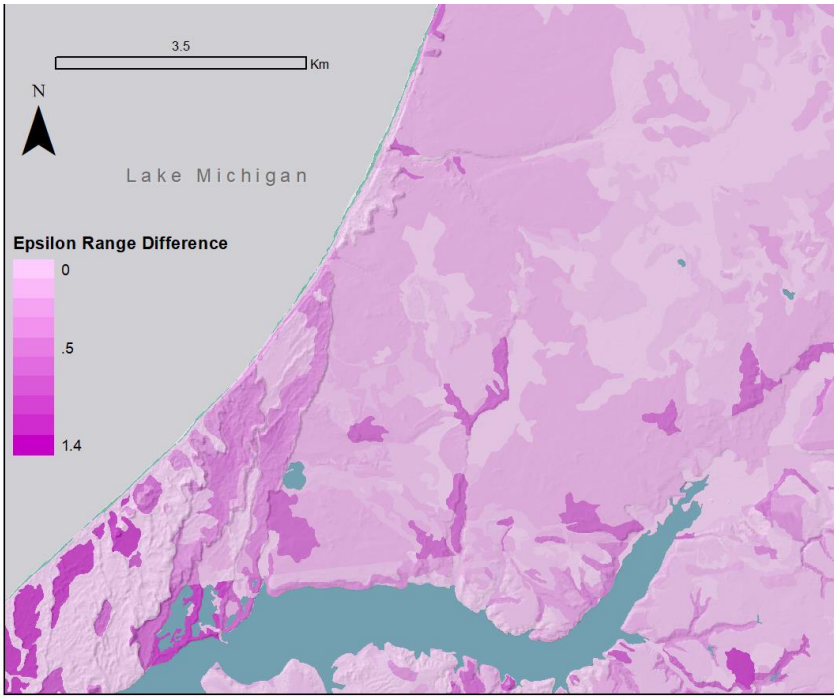


Figure 11 The difference between epsilon minimum and maximum for the Mason County inset. This reflects how difficult it was to assign the soils to the landscape based on the SSURGO descriptions. The higher the value the more manipulation was needed to assign the correct amount of soil.

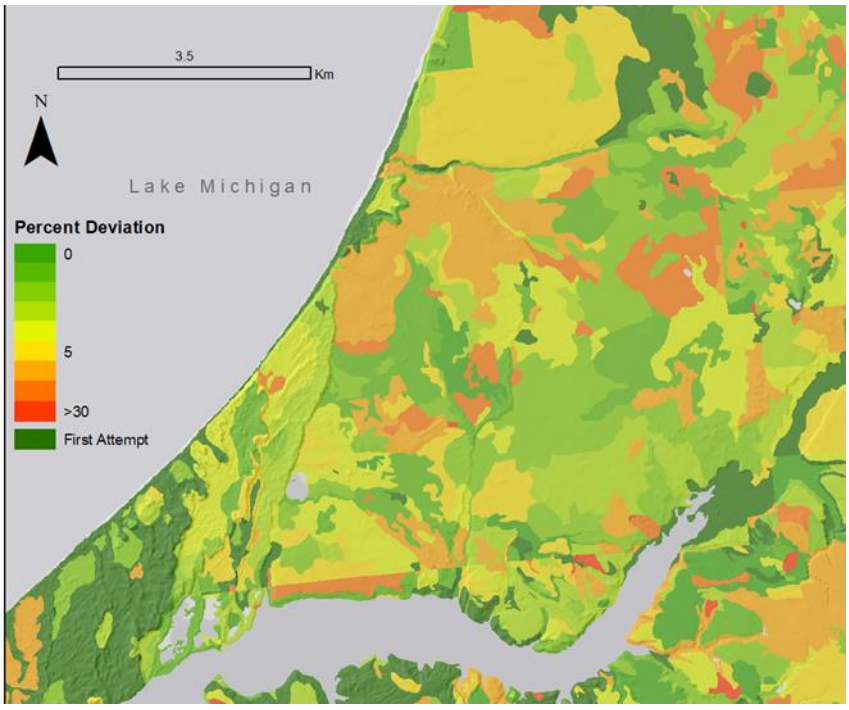


Figure 12 The difference between the area percentage assigned and the area percentage described within the SSURGO data for the Mason County Inset. Higher/redder values reflect further deviation from the SSURGO percentages. Dark green reflects areas where the models assigned the soils on the first attempt.

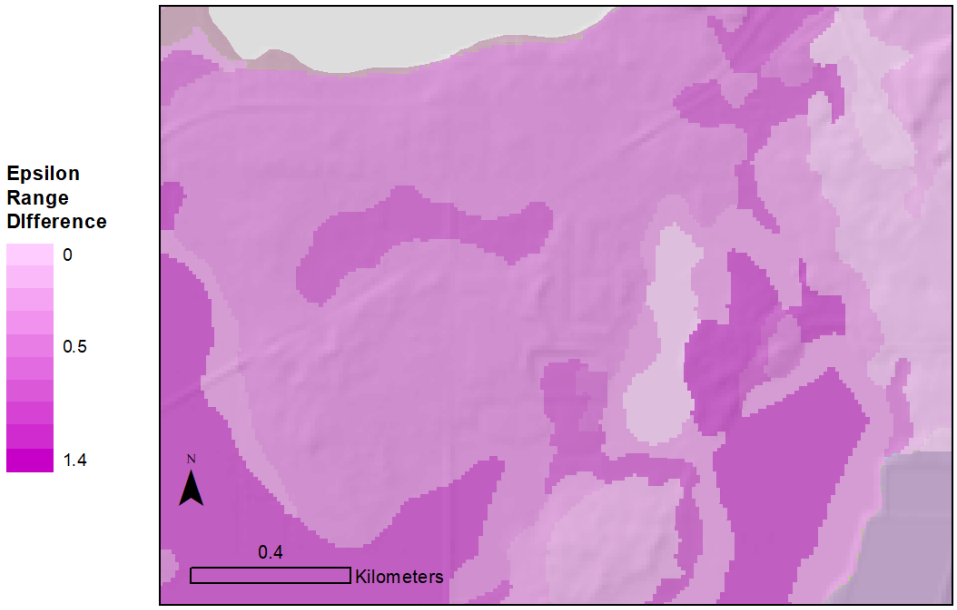


Figure 13 The difference between epsilon minimum and maximum for the Mille Lacs inset. This reflects how difficult it was to assign the soils to the landscape based on the SSURGO descriptions. The higher the value the more manipulation was needed to assign the correct amount of soil.

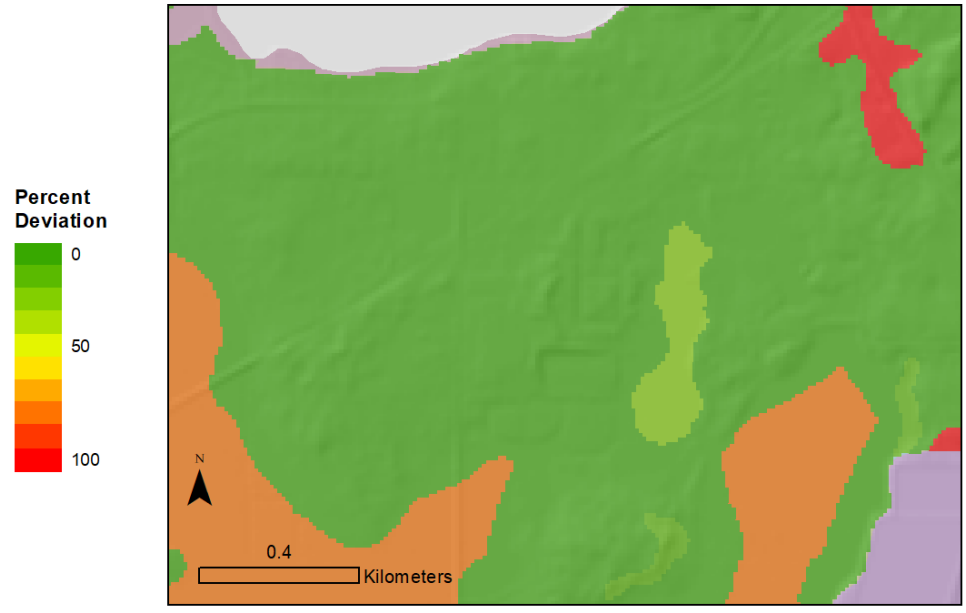


Figure 14 The difference between the area percentage assigned and the area percentage described within the SSURGO data for the Mille Lacs Inset. Higher/redder values reflect further deviation from the SSURGO percentages.

Accuracy Assessment

Finally, the set of evaluation pedon points were compared to both the original soil maps as well as the newly disaggregated maps (Table 8). In the case of Mason County, using all available observation points (n = 167), the original map correctly predicted the correct soil 42 % of the time overall while the disaggregated map correctly predicted the soil 40 % of the time. When using only the randomly sampled observations (n = 41), the SSURGO map marginally underperformed with 41.7 % while the disaggregated map scored 41.9 %. The numbers for the Mille Lacs regions (n =400) were substantially smaller with the SSURGO maps scoring 18% and the disaggregated map scoring 16%.

Separately, accuracies were calculated for disaggregated maps with different sets of values for α, β, γ, and δ (Table **8**). Each alternative set of parameters raised one parameter to 70% and lowered the other parameters to 10%. The increased importance for derivatives calculated from unaltered DEMs (α = 70%) results in best accuracies, with the focal mean DEM variants (β = 70% or δ = 70%) scoring the worst. More extreme values of α, β, γ, and δ had little additional impact.

All accuracies were also calculated using a 3 x 3 window and are reported in Table **8**. The 3 x 3 window improved accuracies of both the SSURGO and the disaggregated maps. However, the increase was generally larger in the disaggregated maps. Looking only at the random sample in Mason County, when using a 3 x 3 window the unaltered DEMs (α = 70%) score increased from 43% to 55% while the original SSURGO map's score only increased from 42% to 46%.

To inspect accuracies in more detail confusion matrices at the level of soil orders were created for the original SSURGO map and the disaggregated map. Additional confusion matrix derivatives, producer and user accuracies, were also calculated (Figure 15, Figure 16, and Table

7). Overall, the model had the most difficulty in distinguishing the Spodosols from the Alfisols in Mason County. Additionally, the model was unable to accurately assign Mollisols within Mason County with a spread of predictions across all soil orders. For the Mille Lacs region, the model had the most difficulty distinguishing the Alfisols from Inceptisols and Mollisols.

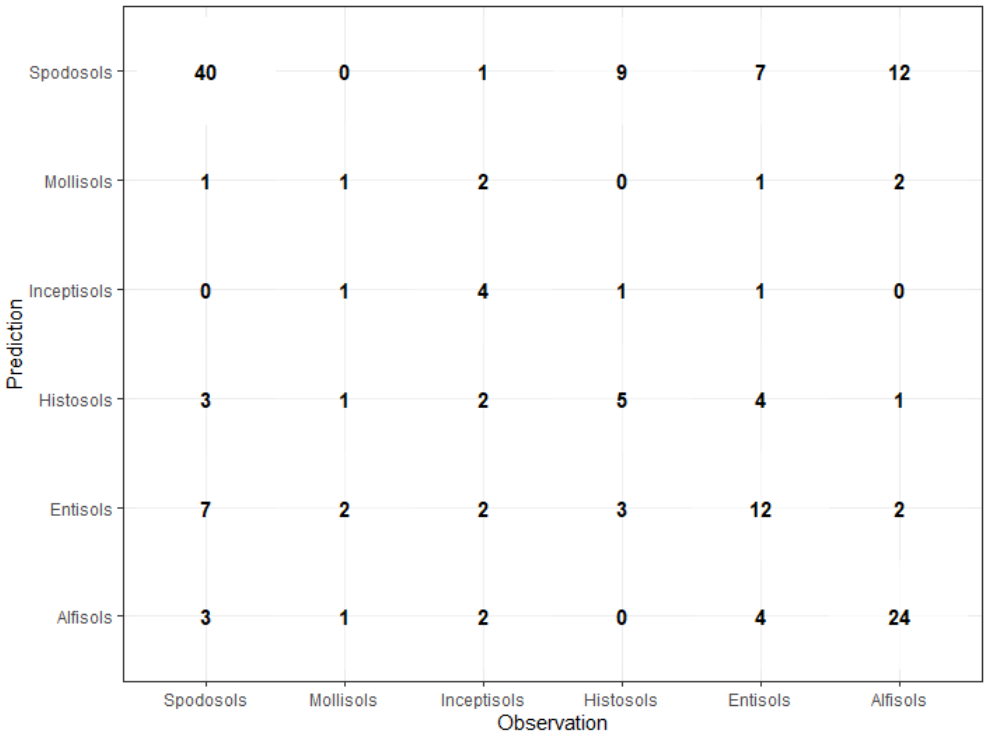


Figure 15 Confusion matrix of soil orders of Mason County with the prediction, the disaggregated map and the reference, the soil observation points.

.

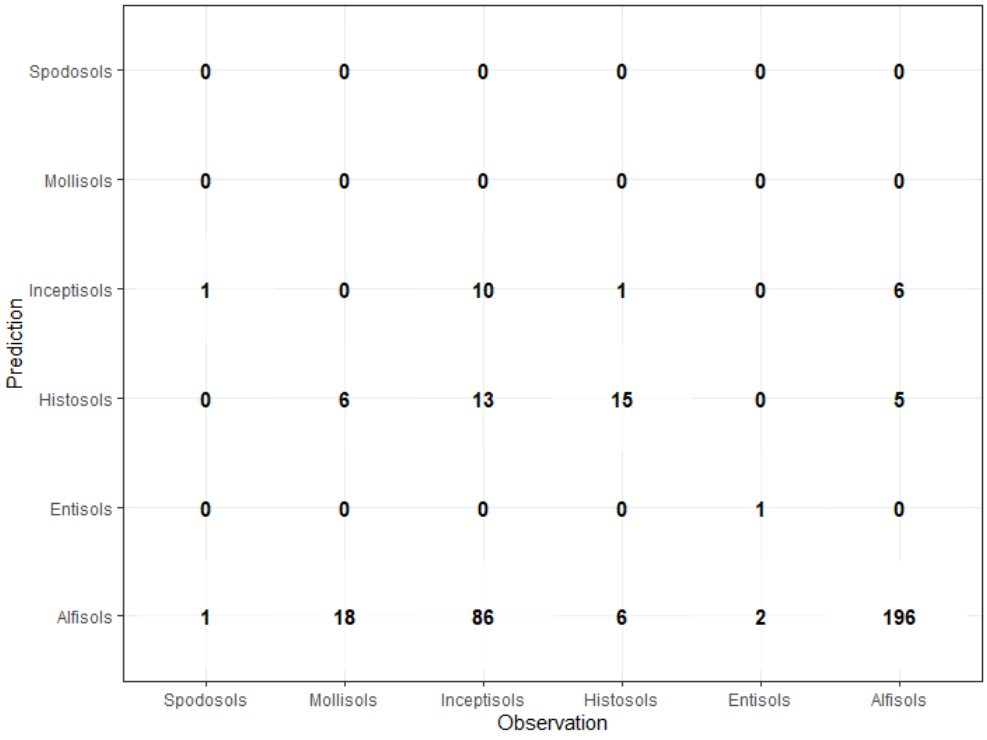


Figure 16 Confusion matrix of soil orders of Mille Lacs region with the prediction, the disaggregated map and the reference, the soil observation points.

Table 7 User's and Producer's accuracies for Mason County and Mille Lacs Region.

Mason County	Soil Order	User's Accuracy	Producer's Accuracy
	Alfisols	71%	59%
	Entisols	43%	41%
	Histosols	31%	28%
	Inceptisols	57%	31%
	Mollisols	14%	17%
	Spodosols	58%	74%
Mille Lacs Region			
	Alfisols	95%	63%
	Entisols	33%	100%
	Histosols	68%	38%
	Inceptisols	9%	56%
	Mollisols	0%	NA
	Spodosols	0%	NA

Table 8 The scores for the original SSURGO maps, the disaggregated maps, and all the variants of the new maps. Overall reflects the score of both the paper copies of NASIS points and the points collected from a random field sample. NASIS is only the paper NASIS soil descriptions and random sample is only the random points collected in the field. In the case of the Mille Lacs region only a set of random NASIS points was used.

Study Site	Observations	Original SSURGO SCORE	25% Equal Weight DEMS	70% Unaltered DEMS	70 % Focal Mean DEMS	70% Aggregated DEMs	70% Focal Mean and
Mason County	All (n=167)	**42%**	40%	41%	38%	40%	37%
	NASIS (n=126)	**42%**	39%	41%	37%	38%	35%
	Random (n=41)	**42%**	42%	43%	39%	47%	44%
Mille Lacs	All (largely random, n=400))	**18%**	16%	19%	18%	17%	16%
3 X 3 Window							
Mason County	All (n=167)	**44%**	46%	47%	46%	44%	45%
	NASIS (n=126)	**44%**	45%	45%	45%	42%	42%
	Random (n=41)	**46%**	49%	55%	48%	50%	54%
Mille Lacs	All (largely random, n=400))	**22%**	24%	26%	25%	25%	22%

Discussion

Practicality of the Approach

Looking at the runtimes and the file sizes it is clear there is a size limit to the practicality of this approach (Table 6 and Table 8). County sized areas seem to work best with having workable runtimes, manageable file sizes, and stability while executing. The biggest limiter to the size of the area that can be disaggregated is the Terrain Index. This index created large files and decreased the stability of the script. This is likely because the Terrain Index was too large to run in memory while the script was running and had to use disk space. The most practical method to avoid this problem was splitting the Mille Lacs region into smaller more manageable sections. Overall, the method is quick and effective for small to county size areas but struggles with larger areas on a traditional office computer. Possible improvements that may be explored include polygon by polygon disaggregation and storing double precision numbers as single precision.

Disaggregated Maps

Looking at the maps themselves it is obvious the number of individual spatial components (i.e. polygons) increased (Figure 8 and Figure 9). Examining further, within the histograms, there was a substantial increase in the number of smaller polygons. This increase reflects the higher resolution of the disaggregated map compared to the original map. Additionally, the increase is seen across the first couple bins within each histogram showing that the polygons increase is not merely the result of single cell "polygons" being created. Overall, this indicates disaggregation successfully assigned more narrowly defined soil map units to more specific locations and the catena information in SSURGO is at least partially reflecting real continuous landforms

Other metrics such as the difference in epsilon ranges and the area percent deviations highlight interesting phenomena within the model and SSURGO data. Overall, areas with higher epsilon differences reflect regions that had to be heavily manipulated to match SSURGO descriptions and assign soils to the landscape. Additionally, a large difference between minimum and maximum epsilon coincided generally with areas with low area percent deviations, indicating methodological success. This is best seen the southwest corner of Figure 11 and Figure 12 where the epsilon values are high and the percent deviation is low. This reflects areas where the model had to work harder to match the landscape to the SSURGO descriptions and provides insight into what areas may need more soil information or more detailed descriptions.

Other possible combinations exist within the maps such as areas with high epsilons and relatively high percent deviations and areas with low epsilons and high percent deviations. In the case of high epsilons and high percent deviations this could be the result of several complications within the model. In these cases, it is likely that the SSURGO description does not match the landscape and no amount of adjustment can overcome and assign the prescribed percent.

Other areas have lower values of epsilon and higher values of deviation from the prescribed percent. In these cases, it is likely because the terrain difference between two components is very small and at a relatively landscape extreme such as 39 % versus 40% slope. As a result when epsilon adjusts only a few pixels may change between each iteration. Since epsilon changes based on area percentages epsilon is put into continuous cycle of alternating percent deviations and as a result the epsilon values settled at the best result they could assign.

Within this study the original component percentages provided in SSURGO were used because they provided the only available target. However, sometimes the slope or curvature of landscape does not allow for the percentages to be met and results in a large deviation from the

SSURGO area percentages. It is possible that adding more terrain information such as aspect, landforms, or raw elevation could reduce this issue by providing additional targets. Additionally, since the SSURGO percentages themselves are estimates it is possible that a lower or higher amount of a soil series is present than what is listed within SSURGO.

Accuracy Assessment

First, looking at the evaluation data for Mason County, the disaggregated maps produce results that are very similar to the original SSURGO data (Table 8). When all points are considered (n = 167) the scores are marginally underperforming SSURGO across all variants of α, β, γ, and δ. SSURGO scores 42% overall with the highest disaggregated map variant (α=70%) scoring 41%.

However, when only considering observations from the random sample (n=41) the disaggregated maps performance varies compared to SSURGO. Almost all the disaggregated map variants match or outperform SSURGO except for (β =70%). This truly random sample is a better test of the maps accuracies because NASIS points are not truly random. NASIS points are often strategically chosen point locations to help create or update soil maps. Additionally, the NASIS points are more random in relation to the disaggregated map than to the SSURGO map.

When the accuracy assessment is done using a 3 X 3 window within Mason County all variants of the disaggregated map match or outperform SSURGO when all points are considered (n =167The only times SSURGO could perform better in a 3 x 3 windows are when the surrounding map units are within the 9-cell window. This suggests some potential boundary issues within the data.

Mille Lacs, even though the scores overall were lower still followed a similar pattern. However, the unaltered DEMS (α = 70 %) marginally outperformed SSURGO without the 3 X 3

window. When expanded to a 3 X 3 window all DEM variants match or outperform SSURGO showing the same pattern to what is seen in Mason County. The lower scores in the Mille Lacs regions are likely the result of the mapping detail and overall information. Overall, only 38.4 % is mapped as order 2 while the entirety of Mason County is mapped as order 2. Additionally, the Mille Lacs disaggregation index consist mostly of singular components. At first this appears as evidence that there is more specific information and that all the components can be isolated. However, the lack of grouped components is the result of more generic descriptions that do not start with many components and as a result are less likely or able to match the sample observation points.

Looking closer overall, between both study sites, at the different variants of α, β, γ, and δ it appears that the unaltered DEMs (α = 70 %) perform the best and approach or surpass SSURGO accuracy. It appears altering the DEMs to change the resulting map shape negatively impacts accuracy in almost all cases. However, the range of accuracies overall demonstrates that weighting does measurably impact the result.

The confusion matrices show some insight into the errors for each region as well. Within Mason County the model had the most difficult time distinguishing the Alfisols from the Spodosols. This is not surprising because Alfisols are soils that leach clay into the subsurface while Spodosols are soils that leach aluminum into the subsurface with both happening in similar, well drained landscape positions. The major classification between them is impossible to determine using landscape position alone. Within the Mille Lacs regions Alfisols had a higher producer's accuracy of 95% but a middling User's accuracy of 63%. This reflects that out of all the Alfisols sample by the reference points 95 % were classified as Alfisols. However, a lot of other points, especially Inceptisols, were also incorrectly classified as Alfisols resulting in the

low User's accuracy. This likely occurred because Inceptisols is a relatively loose soil order of soils that did not meet the criteria of the other soil orders and do not have distinct diagnostic characteristics. Overall, the confusion matrices demonstrate some potential gaps within the method of soils that differ between each other by soil forming factors other than landscape position.

Compared to other similar studies such as POLARIS, (Chaney et al., 2016), this method outperforms POLARIS in soil series level predictions. The most probable soil series within POLARIS at a 30-meter resolution correctly matched the soil series 17% of the time at rank one, 55% at ranks through 10, and 68% when expanded to rank 50.

Within this study only one soil series rank is predicted, and the highest success rate was 41% overall within Mason County at a 10-meter resolution. When expanded to a 3 x 3 search window to match the 30-meter resolution of POLARIS the maximum score further increased to 47%. Overall, the model in this study thus outperforms the POLARIS substantially on a cell-by-cell basis and better matches the deterministic ability of SSURGO to predict the most probable component while sacrificing the national-level validity.

Furthermore, POLARIS by its nature removes artificial discontinuities and boundaries present within the original national-level SSURGO data. Our method disaggregates purely based on SSURGO data and therefore does not remove artificial boundaries.

Conclusion

Overall, the disaggregation procedure produced adequate results. The process appears to work best for smaller areas but can be expanded about the size of the average county without further improvements. Larger areas create bigger and more difficult to manipulate files. The model overall assigned components successfully to much more specific locations and was mostly limited by the descriptions themselves especially in the less detailed mapped Mille Lacs region. The model produced accuracies like that of the SSURGO data and better than that of similar purely digital soil mapping approaches such as POLARIS when determining the soil series present (Chaney et al., 2016). Even though there was not much accuracy improvement from SSURGO the disaggregated maps overall were able to increase the resolution to the SSURGO data while maintaining the overall accuracy of the map, Additionally, regional variations within the SSURGO data are evident by the difference in the accuracies between Mason County and the Mille Lacs region. Existing and disaggregated maps for the Mille Lacs regions underperform those of Mason County. Confusion matrices showed that there is more confusion within Mille Lacs region compared to what is seen in Mason County. These differences are likely a combination of better descriptions and higher order mapping in Mason County compared to the mapping in the Mille Lacs region. Looking ahead, much work still needs to be done to address artificial boundaries and discontinuities within the original maps. Additionally, application of this methodology in other regions of the United States may highlight potential regional issues or benefits to this methodology.

References

Bobryk, C. W., Myers, D. B., Kitchen, N. R., Shanahan, J. F., Sudduth, K. A., Drummond, S. T., Gunzenhauser, B., & Gomez Raboteaux, N. N. (2016). Validating a digital soil map with corn yield data for precision agriculture decision support. *Agronomy Journal*, *108*(3), 957–965. https://doi.org/10.2134/agronj2015.0381

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., & Edwards, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, *239*, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019

Bui, E. N., & Moran, C. J. (2001). Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. *Geoderma*, *103*(1–2), 79–94. https://doi.org/10.1016/S0016-7061(01)00070-2

Causarano, H. J., Doraiswamy, P. C., McCarty, G. W., Hatfield, J. L., Milak, S., & Stern, A. J. (2008). EPIC Modeling of Soil Organic Carbon Sequestration in Croplands of Iowa. *Journal of Environmental Quality*, *37*(4), 1345–1353. https://doi.org/10.2134/jeq2007.0277

Chaney, N. W., Wood, E. F., McBratney, A. B., Hempel, J. W., Nauman, T. W., Brungard, C. W., & Odgers, N. P. (2016). POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. *Geoderma*, *274*, 54–67. https://doi.org/10.1016/j.geoderma.2016.03.025

Davidson, E. A., & Lefebvre, P. A. (1993). Estimating regional carbon stocks and spatially covarying edaphic factors using soil maps at three scales. *Biogeochemistry*, *22*(2), 107–131. https://doi.org/10.1007/BF00002707

FAO-UN - Land and Water Division (CBL). (2007). *Digital Soil Map of the World* (3.6). FAO-UN - Land and Water Division (CBL).

FAO. (2020). *Global Soil Organic Carbon Map (GSOCmap) Version 1.5*. FAO. https://doi.org/10.4060/ca7597en

Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., & Schröder, B. (2012). Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. *Geoderma*, *185–186*, 37–47. https://doi.org/10.1016/j.geoderma.2012.04.001

Hudson, B. D. (1992). The Soil Survey as Paradigm-based Science. *Soil Science Society of America*, *841*(56), 836–841.

ISRIC. (2020). *SoilGrids*.

Janssen, L. L. F., & van der Wel, F. J. M. (1994). Accuracy assessment of satellite derived land-cover data: A review. *Photogrammetric Engineering and Remote Sensing*, *60*(4), 419–426.

Jenny, H. (1941). *The Factors of Soil Formation*. McGraw Hill.

Jiang, H., Radtke, P. J., Weiskittel, A. R., Coulston, J. W., & Guertin, P. J. (2015). Climate- and soil-based models of site productivity in eastern US tree species. *Canadian Journal of Forest Research*, *45*(3), 325–342. https://doi.org/10.1139/cjfr-2014-0054

Jin, Z., Prasad, R., Shriver, J., & Zhuang, Q. (2017). Crop model- and satellite imagery-based recommendation tool for variable rate N fertilizer application for the US Corn system. *Precision Agriculture*, *18*(5), 779–800. https://doi.org/10.1007/s11119-016-9488-z

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. In *Geoderma* (Vol. 117, Issues 1–2). https://doi.org/10.1016/S0016-7061(03)00223-4

Milne, G. (1936). Normal erosion as a factor in soil profile development [9]. In *Nature* (Vol. 138, Issue 3491, pp. 548–549). https://doi.org/10.1038/138548c0

Nachtergaele, F., van Velthuizen, H., & Verelst, L. (2009). *Harmonized World Soil Database*. FAO.

Nauman, T. W., Thompson, J. A., Odgers, N. P., & Libohova, Z. (2012). Fuzzy disaggregation of conventional soil maps using database knowledge extraction to produce soil property maps. *Digital Soil Assessments and Beyond - Proceedings of the Fifth Global Workshop on Digital Soil Mapping, July*, 203–207. https://doi.org/10.1201/b12728-41

Nauman, Travis W., & Thompson, J. A. (2014). Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. *Geoderma*, *213*, 385–399. https://doi.org/10.1016/j.geoderma.2013.08.024

Odgers, N. P., Sun, W., McBratney, A. B., Minasny, B., & Clifford, D. (2014). Disaggregating and harmonising soil map units through resampled classification trees. *Geoderma*, *214–215*, 91–100. https://doi.org/10.1016/j.geoderma.2013.09.024

Simbahan, G. C., Dobermann, A., Goovaerts, P., Ping, J., & Haddix, M. L. (2006). Fine-resolution mapping of soil organic carbon based on multivariate secondary data. *Geoderma*, *132*(3–4), 471–489. https://doi.org/10.1016/j.geoderma.2005.07.001

Soil Survey Staff. (2014). *Keys to Soil Taxonomy* (12th ed.). USDA - Natural Resources Conservation Service.

Soil Survey Staff. (2017). Soil Survey Manual. In *United States Department of Agriculture* (Vol. 18). https://doi.org/10.2307/1233734

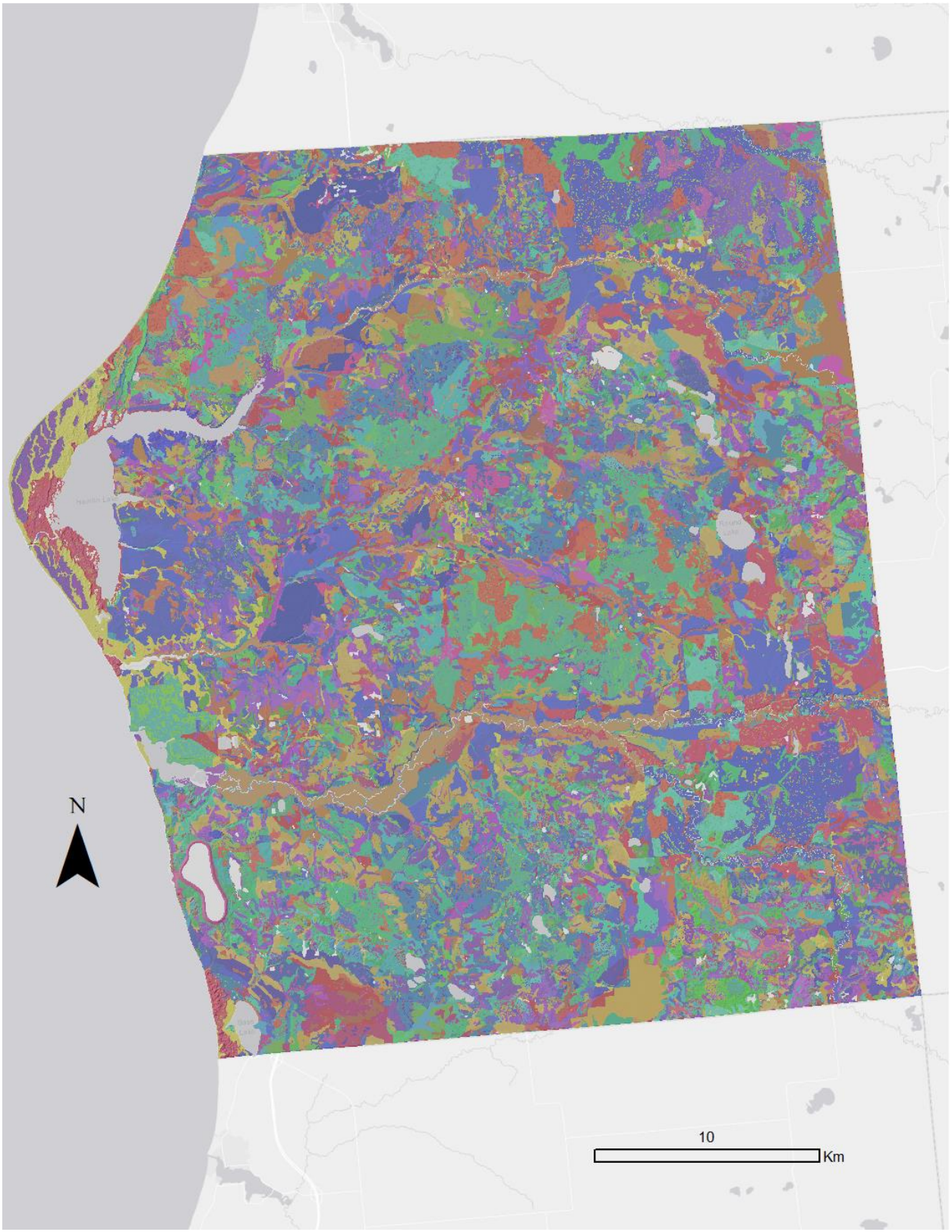
Soil Survey Staff. (2020). *Official Soil Series Description*. Natural Resources Conservation Service, United States Department of Agriculture.

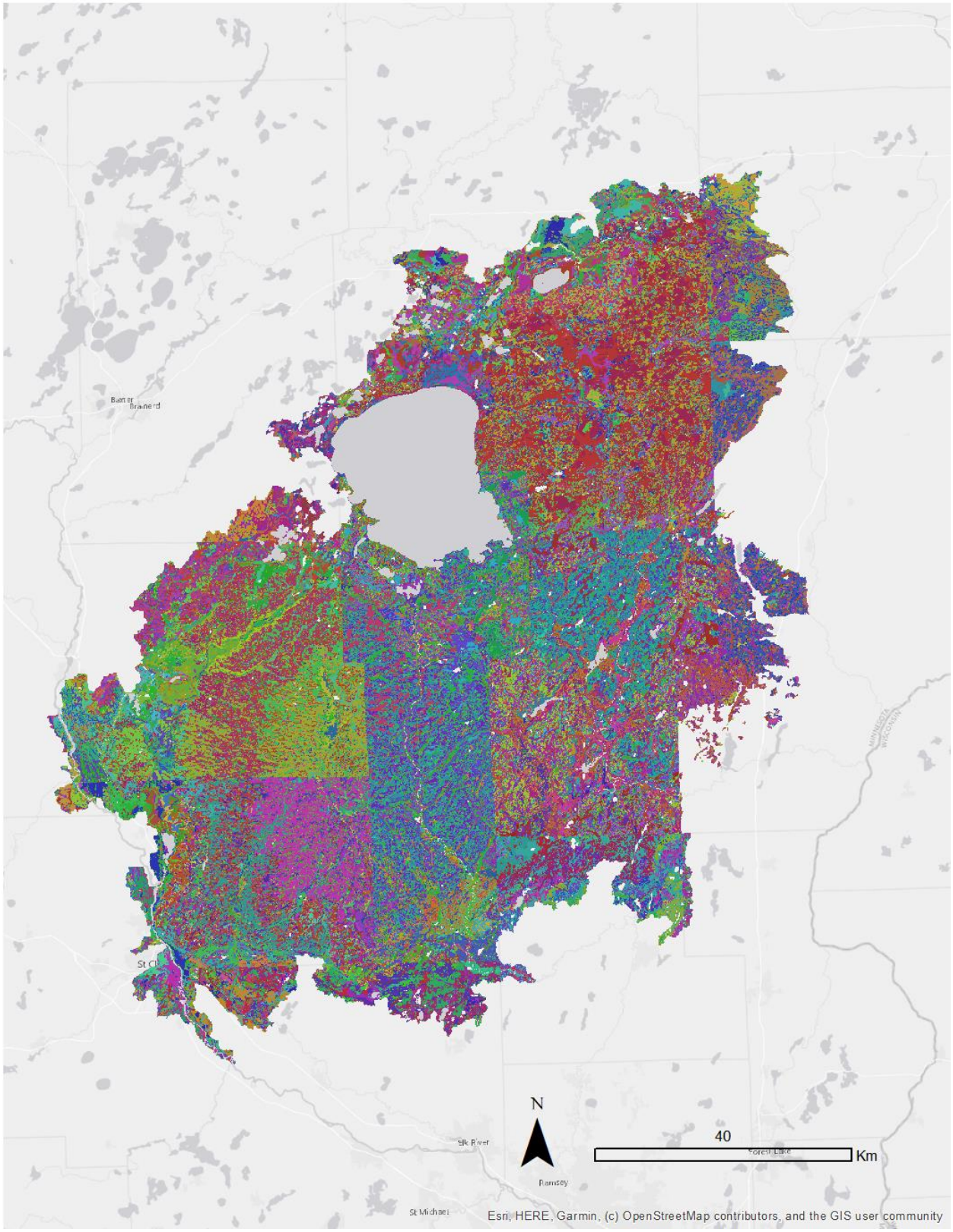
Stoorvogel, J. J., Bakkenes, M., Temme, A. J. A. M., Batjes, N. H., & ten Brink, B. J. E. (2017). S-World: A Global Soil Map for Environmental Modelling. *Land Degradation and Development*, *28*(1), 22–33. https://doi.org/10.1002/ldr.2656

U.S. Geological Survey. (2015). *USGS National Elevation Dataset 1/3 Arc-second*.

Vincent, S., Lemercier, B., Berthier, L., & Walter, C. (2018). Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. *Geoderma, 311*, 130–142. https://doi.org/10.1016/j.geoderma.2016.06.006

Appendix




Brainerd
St Cloud
Elk River
Ramsey
St Michael
Forest Lake
N
40
Km
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community