Analyzing the Use of Publicly Available Multispectral Imagery to Guide the Creation of Soil Sampling Schemes

By

Jordan Paul Oldag

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the degree of Master of Science

Auburn, Alabama
May 2, 2020

Keywords: Remote Sensing, Multispectral Imagery, Geostatistics, Soil Fertility, Soil Sampling

Approved by

Brenda Ortiz, Chair, Professor of Crop, Soil, and Environmental Sciences
Joey Shaw, Professor of Crop, Soil, and Environmental Sciences
Audrey Gamble, Assistant Professor of Crop, Soil, and Environmental Sciences
Rishi Prasad, Assistant Professor of Crop, Soil, and Environmental Sciences
Ruth Kerry, Adjunct Professor of Crop, Soil, and Environmental Sciences

Abstract

As site-specific nutrient management has grown in popularity, the need for accurate soil fertility data has increased. Unfortunately, the cost of detailed soil sampling has prohibited many farmers and consultants from collecting samples at the proper resolution. It is necessary to develop techniques using easy-to-access ancillary data to guide the creation of soil sampling strategies. Ancillary data acquired from three publicly available sources (Landsat 7, Landsat 8, Sentinel 2a) and Soil Electrical Conductivity (EC) was used to determine the strength of relationship to commonly amended soil fertility variables (phosphorus (P), potassium (K), and soil pH). Ancillary data relevancy was determined by a comparison of spatial bi-correlation and fitted semi-variogram ranges. Additionally, principal component analysis (PCA) was performed to allow for unsupervised clustering of ancillary data. These clusters were used to predict zones of nutrient sufficiency. This method resulted in an average class accuracy of 67% (K), 78% (P), and 46% (pH), indicating that classification by clustering may be used to delineate soil fertility distributions to guide the creation of soil sampling schemes.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Adenosine Triphosphate (ATP)
Area of Interest (AOI)
Electrical Conductivity (EC)
Global Positioning System (GPS)
Inverse Distance Weighted (IDW)
Least Squares Support Vector Machine (LS-SVM)
Multivariate Least Squares Regression (MVLSR)
Near-Infrared (NIR)
Partial Least Squares (PLS)
Phosphorus (P)
Potassium (K)
Potential Hydrogen (pH)
Principal Component Analysis (PCA)
Short-Wave-Infrared (SWIR)
Vegetation-Red-Edge (VRE)

CHAPTER I: LITERATURE REVIEW

In the past century, global population has grown from 1.6 to 7.6 billion individuals. This growth in human population has created a need for more sustainable agricultural practices. Despite a population increase of 143.4% between 1961 and 2015, arable land has increased only 25.1%, approximately 976,971,130 hectares (World Bank, 2018). This notable imbalance has caused the arable land, in hectares-per-capita, to be reduced from 0.371 to 0.194. To combat this imbalance, it has been necessary to increase the efficiency of agricultural land usage. In the past century, dramatic improvements have been made in the areas of mechanization, crop breeding, no-tillage implements, pest-management tools, and the use of precision agriculture technology.

Precision agriculture seeks to improve arable land use efficiency by utilizing site-specific management which, in most cases, involves the use of variable rate application of inputs. Until the 1980s, most of these site-specific management strategies were conducted at a farm-level with the field as the smallest management unit (Oliver, 2010). At best, the soil of a field was bulk sampled to determine the mean value of fertility variables and pH, and the field was amended with a uniform application of soil amendments. Since that time, innovations in information technology, sensors, and controls have contributed to more precise management of soil and crop variability (Schueller, 1997). One such innovation is the development of the microprocessor which made it possible for equipment and machines to be outfitted with on-board computers. This allowed for the fitting of global positioning systems (GPS), and sensors for the collection of spatial data for agricultural geographic record keeping enabling common farm equipment to be used for precision agriculture (Robert et al., 1999). In 2019, the International Society of Precision Agriculture defined precision agriculture as, “Precision Agriculture is a management strategy that gathers, processes and

analyzes temporal, spatial and individual data and combines it with other information to support management decisions according to estimated variability for improved resource use efficiency, productivity, quality, profitability and sustainability of agricultural production." To accomplish proper site-specific management, Mulla (2015) defined the four pillars of precision agriculture as: 1) Right source; proper crop variety, fertilizer, herbicide, irrigation source, etc. 2) Right place; to maximize the effectiveness of each source it must be applied in the needed locations, corresponding with within-field variability. 3) Right rate; beyond simply delineating zones of variability, inputs must be applied at variable rates to eliminate waste and increase output. 4) Right time; lastly, these management practices must be implemented at not only a spatially variable rate, but also with temporal variability to increase peak efficiency.

With precision agriculture, inputs, energy, and crops that would be underutilized by traditional management methods can be tailored to best fit the natural characteristics of the land. These site-specific methods seek to eliminate both excess and deficiencies in seeding rates, fertilization, and irrigation, thus reducing excessive or insufficient inputs while maximizing yield. Many of these precision agriculture strategies depend upon a reliable spatial assessment of soil and/or crop distribution. Before decisions can be made about the seeding rates, nutrition, or irrigation accurate soil data acquisition is imperative. Accurate data is required for the detection of varying spatial distribution of nutrients. This variation must be recognized to avoid zones of deficient or excessive application. Depending on the nutrient, both of these application errors can result in reduced nutrient uptake by crops and a decreased yield. Initial soil analysis typically includes data about the texture, fertility, and slope of the arable land.

PHOSPHORUS

Phosphorus (P) is generally considered the second most important macronutrient in agricultural systems production (after nitrogen), typically making up 0.1% to 0.5% of a plant's dry weight (International Plant Nutrition Institute, 1999). Soil P plays many key physiological roles within plants such as formation of nucleic acids, phospholipids, and adenosine triphosphate (ATP). Inadequate levels of available P can result in a reduction in leaf expansion and leaf surface area. Additionally, insufficient P slows the metabolization of carbohydrates in the plant. Without P, the continued production of carbohydrates often results in a dark green or purple color. Because of its importance to cellular and energy production, plants cannot grow without a sufficient P supply. It is therefore necessary to maintain a proper level of P in soils used for cropping systems. The recommended soil extractable threshold for 100% potential yield in Alabama crops is 50 lb./A in sandy soils and loams and 30 lb./A in clayey soils. These recommendations are the same for corn, soybeans, and cotton according to the Alabama Agricultural Experiment Station (Batchelor, 2012). Any amount of extractable P over this threshold is considered to be an unnecessary cost while any deficiencies have potential to adversely affect crop growth and final yield.

It is not uncommon for the total amount of P in soil to be high, but it is often found in plant unavailable forms or outside of the rhizosphere (Schachtman et al., 1998). The relative unavailability of soil P often causes a demand for soil P amendment as few soils release P quickly enough to support optimal growth rates in crop plant species. Proper application of P for agricultural systems requires that farmers must first conduct a proper analysis of the soil-test P, a measure of the P (in solution) available to crops (Knighton and James, 1985). Results from long-term studies across Iowa (approximately 30 years) have shown that soil P trends are greatly affected by the initial P value, cropping years, and annual amendment rates (Dodd and Mallarino, 2005). This indicates the imperative nature of accurate within field assessments of P availability

and variability to guide the total application of P as well as the site-specific variable rate application of this macronutrient.

POTASSIUM

Although potassium (K) is an essential macronutrient for plant growth, it typically has very low concentration in soils ranging from 0.04% to 3% (Sparks and Huang, 1985). Despite its relative rarity, plants accumulate large quantities of this element, which makes up 2% to 10% of a plant's dry weight (Leigh and Jones, 1984; Tisdale et al., 1993). Soil K is vital to many plant processes but is most well-known for its role in enzyme activation. The amount of K in the plant tissue determines the number of enzymes that can be activated and therefore, it influences the rate at which the chemical reactions can occur (Evans and Sorger, 1966; Armstrong, 1987). Soil K is therefore effective at increasing crop yield by enhancing enzyme reactions responsible for root growth, photosynthesis, and transpiration. In addition to this, K produces osmotic pressure helping to regulate the absorption and transportation of water and nutrients.

Soil K does not experience excessive leeching or immobilization and it is typically removed by plant uptake. Because of this, K can often be found in excess in systems where crops do not require the full amount of available K. Due to the low mobility of K in soils, pre-planting soil test is recommended to make future K applications cost-effective and to reduce the risk of over application. For corn and soybeans grown in Alabama cropping systems, it is recommended that soil extractable K is 80 lb./A in sandy soils and loams, and 160 lb./A in clayey soils. Due to its greater need for K, cotton is recommended only in sandy soils with 120 lb./A, loams with 180 lb./A, or clayey soils with 240 lb./A (Batchelor, 2012). A deficiency of plant available K could result in poor crop development or reduced yield. Over application of K has been found to suppress

the uptake of magnesium and calcium (Walsh and O'Donohoe, 1945; Jakobsen, 1993). Accurate assessments of field K can help guide proper application of amendments, avoiding reduced yield from underapplication and unnecessary costs from overapplication.

SOIL pH

Soil potential hydrogen (pH) is a measure of hydrogen ion concentration. Soil pH is typically controlled by clay minerals, organic matter, and oxidation of Al, Fe, Ca, and Na within soils (Thomas, 2006). It is considered by many to be the principal soil variable of agricultural systems (Chakraborty, 2015; Rengel, 2002). Although pH is not a crop nutrient, it is of essential importance to plant growth and crop development. Studies show that excessively acidic soils (low pH) can hamper calcium and magnesium nutrient uptake due to the soluble form of Al^{3+} that results (Tang and Rengel, 2003; Fageria and Zimmermann, 1998). Soil acidity can also affect plant growth by creating toxic levels of soil nutrients or by influencing microbial activity (Mallarino et al., 2011). On the other hand, excessively high pH is known to increase the possibility of herbicide carryover and inhibit plant uptake of soil P, magnesium, iron, manganese, and zinc (Rogovska et al., 2007; Martin and Green, 1989; Lund, 2008). Nutrient unavailability, nutrient toxicity, and harmful microbial activity can each inhibit plant development and yield (Shubert et al, 1990; Fageria and Zimmermann, 1998).

To avoid these problems, it is common practice to apply lime (or another form of calcium carbonate, CaCO3) to increase the soil pH and increase crop production. Unfortunately, sampled pH values alone are not adequate to determine lime recommendations for proper soil amendment (Thomas, 2006). Instead, the buffering capacity of the soil solution, a measure of its ability to resist changes to pH, must be calculated. The soil pH buffer must be determined to estimate the timing,

location, and quantity of lime needed to ensure optimal crop production (Tang and Rengel, 2003). Alabama Agricultural Research Station recommends a soil pH of 6.5 for commercially grown crops (Batchelor, 2012). While uniform liming is the least labor-intensive option, precise application of liming material allows for proper pH balancing and reduces the chance of over-liming (raising the soil pH beyond healthy levels). Proper soil sampling is imperative to determine within-field variability of soil pH and soil pH buffer, in order to accomplish the most efficient application of pH amendments (Lund, 2008).

SOIL SAMPLING

When deciding a sampling strategy for the assessment of soil nutrient levels for soil fertility purposes, it is important to select the strategy that accurately represents the within-field variability of soil nutrient levels. Farmers must evaluate their fertilization goal, soil sampling cost and time, as well as the quality of different sampling strategies in order to use the method best suited for the location.

Grid sampling is done by creating a square or rectangular grid of some defined resolution (such as an area of 2.5 acres) over the area of interest (AOI). This method adopts a uniform approach of sampling across the AOI without accounting for changes in spatial variability of soil nutrients. Because of this, grid sampling is often superior at the detection of small-scale variability assuming that the grid resolution is fine enough to detect small-scale changes in soil properties (Mallarino and Witty, 2001). Unfortunately, this kind of intensive sampling can be cost-prohibitive due to the increased labor costs and soil testing costs (Wollenhaupt and Wolkowski, 1994).

Zone sampling is another popular method, due to its sampling efficiency and reduced costs. This flexible technique uses previous information, sampling objectives, and economic

conditions to provide a customizable method of soil sampling (Saleque et al., 2007). For example, homogenous zones within the AOI may be designed by using soil electrical conductivity (EC) to group the soils in a field into distinct regions. Each zone is assumed to be uniform within its boundaries and therefore requires relatively few samples to create a prescription maps of soil properties by zone (Shaner et al., 2008). Zones can be created from historical soil data, ancillary data, or simply by farmer observation. Unfortunately, when using zone sampling, the quality of results changes from field to field. Consequently, site-specific decisions must be made to determine if grid sampling is more appropriate for an AOI or if zones can be delineated that properly capture the spatial variation of soil nutrients. The majority of precision agriculture operations sample using a grid sampling strategy. This method has grown in popularity due to its relatively uniform approach at soil nutrient detection and its cost-effective nature (McBratney and Pringle, 1999; Robert et al., 1999; Winstead et al., 2007).

No matter which sampling strategy is chosen, it is necessary that the sampling scheme created accurately captures the spatial variability of each soil nutrient. Samples are collected to be used in the creation of a prescription map for the variable rate application of soil amendments. It is necessary that the sampling strategy used produces a prescription map with high accuracy, as any errors will result in unnecessary inputs or insufficient soil fertilization. Without proper detection of soil nutrient variability accurate estimations of soil nutrients cannot be created and variable rate application cannot be accurately used.

To accomplish these accurate predictions, a minimum of approximately 100 data points is required the creation of a reliable variogram, which may be used for the spatial interpolation process known as kriging (Webster and Oliver, 1992). This method is often used incorrectly, as it requires a much larger sample-size than most commercial farms collect. Most crop production soil

surveys are conducted at the minimum recommended resolution of a 2.5-acre (1 hectare or 100 m) grid (Ferguson and Hergert, 2000; Godwin and Miller, 2003). This sampling technique results in too small of a sample for most fields and results in the creation and use of unstable variograms for soil predictions. The arbitrary use of the 2.5-acre sampling interval often fails to capture the scale of variation present within the field. It is essential that the sampling interval selected captures data that is spatially dependent and correlated to allow for kriging and mapping of soil variables (Kerry and Oliver, 2008; Lopez-Granados et al., 2002).

GEOSTATISTICS

Geostatistics is a class of statistics used for the description, interpretation, and estimation of data which is associated with spatial or spatiotemporal phenomena (Wackernagel, 1995). It is therefore necessary to have spatial features (coordinates) associated with the observations being analyzed. Least squares and linear estimation methods were initially adapted for the mining industry for the estimation of quantities that vary in space (Kitanidis, 1997). To do this, geostatistics investigates spatial autocorrelation which can be summarized using Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). The semi-variogram $\gamma(h)$ is the central tool of geostatistics, which was defined by Matheron (1963) as half the average squared difference between points separated at distance h. The empirical semi-variogram is calculated as:

$$\gamma(h) = \frac{1}{2|N(h)|} \sum_{N(h)} (z_i - z_j)^2$$

where *N(h)* is the set of all pairwise Euclidean distances $i - j = h$, $|N(h)|$ is the number of distinct pairs in *N(h)*, and z_i *and* z_j are data values at spatial location *i* and *j*, respectively. *h* represents a distance measure with magnitude only (Ford, 2010).

Typically, semi-variance observations are grouped into bins by lag size selection, $h \pm \delta$ rather than using exact distances. Therefore, the empirical semi-variograms $\hat{\gamma}(h \pm \delta)$ can be calculated for each bin:

$$\hat{\gamma}(h \pm \delta) = \frac{1}{2|N(h \pm \delta)|} \sum_{(i,j) \epsilon\, N(h \pm \delta)} |z_i - z_j|^2$$

Each pair of points separated by *h* plus or minus lag size range δ are found. These form the set of points $N(h \pm \delta) \equiv \{(s_i, s_j) : |s_i, s_j| = h \pm \delta; i, j = 1, \ldots, N\}$. The number of points in this bin is $N(h \pm \delta)$. For each pair of points *i, j*, the square of the difference in the observation is found $(|z_i - z_j|)^2$. These squared differences are added together and normalized by the natural number$|N(h \pm \delta)|$ before being divided by two, to yield the semi-variogram. The measure of variability between pairs of points in each lag, up to the maximum specified lag, are then used to create the experimental semi-variogram as a discrete function. A model can then be fitted to the experimental semi-variogram resulting in a continuous function that can be used in kriging, the interpolation process (Leyk, 2008).

When approaching the subject of interpolation of values at unobserved locations, the value of the semi-variogram to geostatistics becomes apparent. If the data exemplifies spatial autocorrelation, prediction may be appropriate, by use of the fitted model (spherical, exponential, Gaussian, etc.) that best captures the spatially dependent variation of the dataset by minimizing residuals and predictive error. The semi-variogram model can then be used for estimation by

kriging. The process of kriging estimates the value of an unobserved location by interpolating from weighted points within the determined range of the semi-variogram model. To accomplish this, three major parameters are used to guide the kriging process. The range, or the distance at which auto-correlation between points can no longer be observed, is used to determine the suggested search distance for the interpolation process. The nugget is a measure of variation in observations that occurs at distances smaller than the sampling interval. Lastly, the sill represents the variation in observations that occurs at the range, where autocorrelation is no longer observed.

In addition to univariate kriging, allowing predictions to unobserved locations, ancillary data can be used to create cross-variograms, to estimate variables that vary in space with a similar distribution as ancillary variables (Zhang et al., 1992; Hong et al., 2002; Pimstein et al., 2011). These techniques have been used to interpolate permanent soil properties from sparsely sampled sites, by using a sampling interval of half average variogram range ancillary data (Kerry and Oliver, 2003; Kerry and Oliver, 2004).

ANCILLARY DATA

Remotely sensed data has long been considered a valuable source of ancillary data to guide precision agriculture processes (Mulla, 1997). Remotely sensed ancillary data has been used to study the prediction of soil characteristics including soil texture, extractable iron, total carbon (Sullivan et al., 2005), cation exchange capacity, magnesium, organic matter (Hong et al., 2002), and soil hydraulic properties (Mohanty, 2013). The data in a majority of past studies is collected from remotely sensed imagery (aerial imagery, unmanned aerial vehicle imagery, and satellite imagery). These studies tend to return a wide range of results depending on prediction technique

and study location. Each soil type requires unique study and techniques for regression methods to accurately predict soil nutrients (Yu et al., 2016).

Table 1 shows a collection of studies that include the use of ancillary data (multispectral/hyperspectral imagery, soil EC, yield) and their correlation to soil macronutrients P and K and soil pH. Most of these studies are conducted with bare soil reflectance values to investigate the strength of various ancillary data sources. As a whole, these studies indicate that K has the strongest relationship with spectral reflectance data and achieves the most accurate predictions with least square support vector machines (LS-SVM) and multivariate least squares regression (MLSR). Likewise, P was predicted with relative strength using partial least squares (PLS) regression. The relationship between soil pH and remotely sensed ancillary data has not been studied as thoroughly, the most accurate predictions are created by use of soil EC.

For the purposes of this study, ancillary data will not be used for the prediction of soil macronutrients. Ancillary data will instead be used for the prediction of zones corresponding to the spatial scale or soil fertility variables. Variograms created from ancillary data show the range of spatial dependency, which can be used to predict the spatial range of soil variables and determine the sampling interval for soil sampling (Kerry and Oliver, 2003). It is recommended that the site-specific sampling scheme is created at half of the range of spatial dependence observed in the variogram from corresponding ancillary data (Kerry and Oliver, 2004). This approach is chosen to ensure that the sample locations safely capture the variability of soil nutrients. If used to identify zones of predicted high or low soil macronutrient content, the range of spatial dependence observed in ancillary data can be used to create site specific soil sampling schemes.

Table 1. Soil Nutrient Prediction Review

Predicted Variable(s)	Ancillary Variable(s)	Materials and Methods	Study
Ca, Mg, P, K	Soil EC	Relationships determined by PCA in individual fields. (R^2=0.66 - 0.93)	Heiniger et al. (2003)
Soil EC (R^2=0.56), CEC (R^2=0.66), pH (R^2=0.68), Mg (0.67), K (R^2=0.59)	Hyperspectral Imagery (soil)	SMLR Models using all bands.	Hong et al. (2002)
Yield	Soil EC	Simple regression to create nitrogen and lime prescription.	Lund et al. (1999)
N (0.31 - 0.38), P (0.31 - 0.42), S (0.42 - 0.53), K (0.29 - 0.36)	Hyperspectral Imagery (vegetation)	Regressive models of various vegetative indices.	Mahajan et al. (2014)
Phosphorus deficiency	Spectral Reflectance (0.68 um)	Shift of red edge to shorter wavelengths.	Milton et al. (1991)
pH (0.66), CEC (0.843)	Soil EC	Correlation given from estimated soil Coarse-sand Content and Clay Content from mobile Soil EC.	Moral et al. (2010)
P (R2 0.61 - 0.68), N	Hyperspectral Reflectance (vegetation)	Prediction of P deficiencies present in blue and NIR through early growing season.	Osborne et al (2002)
P (DS1 + 2, 0.68), K (DS2, 0.88)	Spectral Reflectance	Prediction models created with subset data for each nutrient using PLS.	Pimstein et al. (2011)
K (0.6858), Soil OM (0.8708), N (0.7206)	Spectral Reflectance	Prediction created by PCA and least square support vector machine (LS-SVM)	Qiao and Zhang (2012)

K (0.95 - 0.97)	Spectral Reflectance (Soil)	MLSR from Productive Index (PI)	Yu et al. (2016)
K, Ca, N Fertilizer	Agronomic Efficiencies and Sustainable Yield Index	Study analyzes yield increase attributed to fertilization recommendations created by AE and SYI.	Liu et al. (2017)
Millet Yield	Soil CEC and Elevation	Linear regression explained approximately 30% of variation observed.	Stein et al. (1997)
Soybean Yield (0.05 – 0.71)	Soil Fertility and Topography	Stepwise multiple regression with varying results depending on yield limiting factor.	Kravchenko and Bullock (2000)
Soil Depth, Volumetric Water Content	Yield, Color Photographs, Soil EC, Stoniness	Comparison of lag distances between permanent soil properties and ancillary data.	Kerry and Oliver (2003)

RESEARCH OBJECTIVES

It is expected that remotely sensed ancillary data can be used to predict soil variable values for the characteristics described above (Kerry and Oliver, 2003; Ben-Dor, 2002; Agbu et al., 1990). This study focused on method development for soil sampling scheme estimation from predicted soil characteristics from ancillary data. Publicly available remotely sensed data was used (Landsat 7, Landsat 8, Sentinel 2a) to monitor temporal and spatial variability of multispectral recordings. Soil sampling with nutrient analysis was also conducted to verify ancillary results expected from literature on the topic.

The ability to create site-specific soil sampling schemes to accurately represent within-field variability of soil variables is critical for popularizing the use of precision agriculture technologies. Site-specific sampling schemes will allow for cost-efficient and profitable collection and use of spatial and temporal variability of soil variables. This study seeks to develop a methodology that best characterizes within-field variability among soil variables (P, K, and pH) in a way that leads to optimum soil sampling. The two objectives of this project are: 1) evaluate the use of ancillary data for indirect assessment of P, K, and pH variability on large scale production fields; and, 2) develop a method to predict soil fertility zones to guide creation of soil sampling schemes.

LITERATURE CITED

Agbu, P.A., D.J. Fehrenbacher, and I.J. Jansen. 1990. Soil Property Relationships with SPOT Satellite Digital Data in East Central Illinois. Soil Science Society of America Journal 54(3): 807–812. doi: 10.2136/sssaj1990.03615995005400030031x.

Armstrong, D., editor. 1987. Functions of Potassium in Plants. Potassium for Agriculture. Potash and Phosphate Institute, Atlanta, GA. p. 4–5

Batchelor, W. 2012. Nutrient Recommendation Tables for Alabama Crops.

Ben-Dor, E. 2002. Quantitative remote sensing of soil properties. Advances in Agronomy. Academic Press. p. 173–243

Chakraborty, K. 2015. Soil pH as a Master Variable of Agricultural Productivity in Burdwan I CD Block.

Dodd, J.R., and A.P. Mallarino. 2005. Soil-Test Phosphorus and Crop Grain Yield Responses to Long-Term Phosphorus Fertilization for Corn-Soybean Rotations. Soil Science Society of America Journal 69(4): 1118–1128. doi: 10.2136/sssaj2004.0279.

Evans, H.J., and G.J. Sorger. 1966. Role of Mineral Elements with Emphasis on the Univalent Cations. Annual Review of Plant Physiology 17(1): 47–76. doi: 10.1146/annurev.pp.17.060166.000403.

Fageria, N.K., and F.J.P. Zimmermann. 1998. Influence of pH on growth and nutrient uptake by crop species in an Oxisol. Communications in Soil Science and Plant Analysis 29(17–18): 2675–2682. doi: 10.1080/00103629809370142.

Ferguson, R.B., and G.W. Hergert. 2000. Soil Sampling for Precision Agriculture. University of Nebraska Extension EC 00-154: 4.

Ford, D. 2010. The Empirical Variogram. http://faculty.washington.edu/edford/Variogram.pdf (accessed 2 October 2018).

Godwin, R.J., and P.C.H. Miller. 2003. A Review of the Technologies for Mapping Within-field Variability. Biosystems Engineering 84(4): 393–407. doi: 10.1016/S1537-5110(02)00283-0.

Goovaerts, P., and R. Kerry. 2010. Using Ancillary Data to Improve Prediction of Soil and Crop Attributes in Precision Agriculture. In: Oliver, M.A., editor, Geostatistical Applications for Precision Agriculture. Springer Netherlands, Dordrecht. p. 167–194

Heiniger, R.W., R.G. McBride, and D.E. Clay. 2003. Using Soil Electrical Conductivity to Improve Nutrient Management. Agronomy Journal 95(3): 508–519. doi: 10.2134/agronj2003.5080.

Hong, S.Y., K.A. Sudduth, N.R. Kitchen, S.T. Drummond, H.L. Palm, et al. 2002. Estimating Within-Field Variations in Soil Properties from Airborne Hyperspectral Images. Conference Proceedings: 13.

International Plant Nutrition Institute. 1999. Functions of Phosphorus in Plants. Better Crops with Plant Food 83(1): 6–7.

International Society of Precision Agriculture. Precision Ag Definition. https://ispag.org/about/definition (accessed 8 December 2019).

Jakobsen, S.T. 1993. Interaction between Plant Nutrients: III. Antagonism between Potassium, Magnesium and Calcium. Acta Agriculturae Scandinavica, Section B — Soil & Plant Science 43(1): 1–5. doi: 10.1080/09064719309410223.

Kerry, R., and M.A. Oliver. 2003. Variograms of Ancillary Data to Aid Sampling for Soil Surveys. Precision Agriculture 4(3): 261–278. doi: 10.1023/A:1024952406744.

Kerry, R., and M.A. Oliver. 2004. Average variograms to guide soil sampling. International Journal of Applied Earth Observation and Geoinformation 5(4): 307–325. doi: 10.1016/j.jag.2004.07.005.

Kerry, R., and M.A. Oliver. 2008. Determining nugget:sill ratios of standardized variograms from aerial photographs to krig sparse soil data. Precision Agriculture 9(1–2): 33–56. doi: 10.1007/s11119-008-9058-0.

Kitanidis, P.K. 1997. Introduction to Geostatistics: Applications in Hydrogeology. Cambridge University Press.

Knighton, R.E., and D.W. James. 1985. Soil Test Phosphorus as a Regionalized Variable in Leveled Land 1. Soil Science Society of America Journal 49(3): 675–679. doi: 10.2136/sssaj1985.03615995004900030030x.

Kravchenko, A.N., and D.G. Bullock. 2000. Correlation of Corn and Soybean Grain Yield with Topography and Soil Properties. Agronomy Journal 92(1): 75–83. doi: 10.2134/agronj2000.92175x.

Lee, W.S., J.F. Sanchez, R.S. Mylavarapu, and J.S. Choe. 2003. ESTIMATING CHEMICAL PROPERTIES OF FLORIDA SOILS USING SPECTRAL REFLECTANCE. Transactions of the ASAE 46(5). doi: 10.13031/2013.15438.

Leigh, R.A., and R.G.W. Jones. 1984. A Hypothesis Relating Critical Potassium Concentrations for Growth to the Distribution and Functions of This Ion in the Plant Cell. New Phytologist 97(1): 1–13. doi: 10.1111/j.1469-8137.1984.tb04103.x.

Leyk, S. 2008. Spatial Estimation. https://www.colorado.edu/geography/class_homepages/geog_4203_s08/class9_SpatialEstimationB.pdf (accessed 2 October 2018).

Liu, C., Y. Liu, Z. Li, G. Zhang, and F. Chen. 2017. A novel way to establish fertilization recommendations based on agronomic efficiency and a sustainable yield index for rice crops. Scientific Reports 7(1): 1001. doi: 10.1038/s41598-017-01143-2.

López-Granados, F., M. Jurado-Expósito, S. Atenciano, A. García-Ferrer, M.S. de la Orden, et al. 2002. Spatial variability of agricultural soil parameters in southern Spain. Plant and Soil 246(1): 97–105. doi: 10.1023/A:1021568415380.

Lund, E.D. 2008. Measuring and Managing Soil pH. Soil Science Step-by-Step Field Analysis. Soil Science Society of America, Madison, WI. p. 147–158

Lund, E.D., P.E. Colin, D. Christy, and P.E. Drummond. 1999. Applying Soil Electrical Conductivity Technology to Precision Agriculture. Precision Agriculture access publication(precisionagric4b): 1089–1100. doi: 10.2134/1999.precisionagproc4.c12b.

Mahajan, G.R., R.N. Sahoo, R.N. Pandey, V.K. Gupta, and D. Kumar. 2014. Using hyperspectral remote sensing techniques to monitor nitrogen, phosphorus, sulphur and potassium in wheat (Triticum aestivum L.). Precision Agriculture 15(5): 499–522. doi: 10.1007/s11119-014-9348-7.

Mallarino, A., A. Pagani, and J. Sawyer. 2011. Corn and soybean response to soil pH level and liming. Integrated Crop Management Conference: 93–102.

Mallarino, A.P., and D. Witty. 2001. Management Zones Sampling: A Better Alternative. Iowa State University.

Martin, J., and J. Green. 1989. Herbicide persistence and carryover. University of Kentucky Extension Bullington AGR-139. http://www2.ca.uky.edu/agcomm/pubs/agr/agr139/agr139.htm (accessed 31 January 2018).

Matheron, G. 1963. Principles of geostatistics. Economic Geology 58(8): 1246–1266. doi: 10.2113/gsecongeo.58.8.1246.

McBratney, A.B., and M.J. Pringle. 1999. Estimating Average and Proportional Variograms of Soil Properties and Their Potential Use in Precision Agriculture. Precision Agriculture 1(2): 125–152. doi: 10.1023/A:1009995404447.

Milton, N.M., B.A. Eiswerth, and C.M. Ager. 1991. Effect of phosphorus deficiency on spectral reflectance and morphology of soybean plants. Remote Sensing of Environment 36(2): 121–127. doi: 10.1016/0034-4257(91)90034-4.

Mohanty, B.P. 2013. Soil Hydraulic Property Estimation Using Remote Sensing: A Review. Vadose Zone Journal 12(4). doi: 10.2136/vzj2013.06.0100.

Mulla, D. J. 1997. Geostatistics, remote sensing and precision farming. In: Precision Agriculture: Spatial and Temporal Variability of Environmental Quality, edited by J. V. Lake, G. R. Bock and J. A. Goode (John Wiley and Sons Ltd., Chichester, UK), pp. 110-119.

Mulla, D.J. 2015. Spatial Variability in Precision Agriculture. In: Shekhar, S., Xiong, H., and Zhou, X., editors, Encyclopedia of GIS. Springer International Publishing, Cham. p. 1–8

Oliver, M.A. 2010. An Overview of Geostatistics and Precision Agriculture. In: Oliver, M.A., editor, Geostatistical Applications for Precision Agriculture. Springer Netherlands, Dordrecht. p. 1–34

Osborne, S.L., J.S. Schepers, D.D. Francis, and M.R. Schlemmer. 2002. Detection of Phosphorus and Nitrogen Deficiencies in Corn Using Spectral Radiance Measurements. Agronomy Journal 94(6): 1215–1221. doi: 10.2134/agronj2002.1215.

Pimstein, A., A. Karnieli, S.K. Bansal, and D.J. Bonfil. 2011. Exploring remotely sensed technologies for monitoring wheat potassium and phosphorus using field spectroscopy. Field Crops Research 121(1): 125–135. doi: 10.1016/j.fcr.2010.12.001.

Qiao, Y., and S. Zhang. 2012. Near-Infrared Spectroscopy Technology for Soil Nutrients Detection Based on LS-SVM. In: Li, D. and Chen, Y., editors, Computer and Computing Technologies in Agriculture V. Springer Berlin Heidelberg. p. 325–335

Rengel, Z. 2002. Handbook of Plant Growth: pH as the Master Variable. Marcel Dekker, Inc, New York, NY.

Robert, P.C., R.H. Rust, W.E. Larson, K.L. Fleming, D.G. Westfall, et al. 1999. Evaluating Farmer Developed Management Zone Maps for Precision Farming. ACSESS publications. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America

Rogovska, N.P., A.M. Blackmer, and A.P. Mallarino. 2007. Relationships between Soybean Yield, Soil pH, and Soil Carbonate Concentration. Soil Science Society of America Journal 71(4): 1251–1256. doi: 10.2136/sssaj2006.0235.

Saleque, M.A., M.K. Uddin, A.K.M. Ferdous, and M.H. Rashid. 2007. Use of Farmers' Empirical Knowledge to Delineate Soil Fertility-Management Zones and Improved Nutrient-Management for Lowland Rice. Communications in Soil Science and Plant Analysis 39(1–2): 25–45. doi: 10.1080/00103620701758915.

Schachtman, D.P., R.J. Reid, and S.M. Ayling. 1998. Phosphorus Uptake by Plants: From Soil to Cell. Plant Physiology 116(2): 447–453. doi: 10.1104/pp.116.2.447.

Schubert, S., E. Schubert, and K. Mengel. 1990. Effect of low pH of the root medium on proton release, growth, and nutrient uptake of field beans (Vicia faba). In: van Beusichem, M.L., editor, Plant Nutrition — Physiology and Applications: Proceedings of the Eleventh International Plant Nutrition Colloquium, 30 July–4 August 1989, Wageningen, The Netherlands. Springer Netherlands, Dordrecht. p. 443–448

Schueller, J.K. 1997. Technology for Precision Agriculture. Precision Agriculture: 19–33.

Shaner, D.L., R. Khosla, M.K. Brodahl, G.W. Buchleiter, and H.J. Farahani. 2008. How Well Does Zone Sampling Based on Soil Electrical Conductivity Maps Represent Soil Variability? Agronomy Journal 100(5): 1472–1480. doi: 10.2134/agronj2008.0060.

Sparks, D.L., and P.M. Huang. 1985. Physical Chemistry of Soil Potassium. Potassium in Agriculture. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI. p. 201–276

Stein, A., J. Brouwer, and J. Bouma. 1997. Methods for Comparing Spatial Variability Patterns of Millet Yield and Soil Data. Soil Science Society of America Journal 61(3): 861–870. doi: 10.2136/sssaj1997.03615995006100030021x.

Sullivan, D.G., J.N. Shaw, and D. Rickman. 2005. IKONOS Imagery to Estimate Surface Soil Property Variability in Two Alabama Physiographies. Soil Science Society of America Journal 69(6): 1789. doi: 10.2136/sssaj2005.0071.

Tang, C., and Z. Rengel. 2003. Role of Plant Cation/Anion Uptake Ratio in Soil Acidification. In: Rengel, Z., editor, Handbook of Soil Acidity. Marcel Dekker, Inc, New York, NY. p. 70

Thomas, G.W. 2006. pH (R. Lal, editor). Encyclopedia of Soil Science.

Tisdale, S., W. Nelson, and J. Beaton. 1993. Soil Fertility and Fertilizers. 5th ed. Pearson.

Tobler, W.R. 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography 46: 234. doi: 10.2307/143141.

Wackernagel, H. 1995. Introduction. Multivariate Geostatistics. p. 1–4

Walsh, T., and T.F. O'Donohoe. 1945. Magnesium deficiency in some crop plants in relation to the level of potassium nutrition. The Journal of Agricultural Science 35(4): 254–263. doi: 10.1017/S0021859600013587.

Webster, R., and M.A. Oliver. 1992. Sample adequately to estimate variograms of soil properties. Journal of Soil Science 43(1): 177–192. doi: 10.1111/j.1365-2389.1992.tb00128.x.

Winstead, A.T., J.N. Shaw, P.L. Mask, and S.H. Norwood. 2007. Precision Soil Sampling for Alabama Farms. Alabama Cooperative Extension System: 3.

Wollenhaupt, N.C., and R.P. Wolkowski. 1994. Grid Soil Sampling. Better Crops 78(4): 6–9.

World Bank. "Arable land (% of land area)". *World Development Indicators*. The World Bank Group, 2015, https://data.worldbank.org/indicator/AG.LND.ARBL.ZS

Yu, X., Q. Liu, Y. Wang, X. Liu, and X. Liu. 2016. Evaluation of MLSR and PLSR for estimating soil element contents using visible/near-infrared spectroscopy in apple

orchards on the Jiaodong peninsula. CATENA 137: 340–349. doi: 10.1016/j.catena.2015.09.024.

Zhang, R., A.W. Warrick, and D.E. Myers. 1992. Improvement of the prediction of soil particle size fractions using spectral properties. Geoderma 52(3–4): 223–234. doi: 10.1016/0016-7061(92)90038-9.

CHAPTER II: USING SPATIAL CORRELATION ANALYSIS TO IDENTIFY ANCILLARY DATA FOR SOIL FERTILITY MAPPING

ABSTRACT

As site-specific nutrient management has grown in popularity, the need for accurate soil fertility data has increased. Unfortunately, the cost of detailed soil sampling prohibits many farmers and consultants from collecting samples at the proper resolution. It is necessary to develop techniques using easy-to-access ancillary data to guide the creation of soil sampling strategies. Ancillary data acquired from three sources Landsat 8, Sentinel 2a, and Soil Electrical Conductivity (EC) was used to determine the strength of relationship to commonly amended soil fertility variables, phosphorus (P), potassium (K), and soil pH. First, strength-of-relationship was determined by Pearson Correlation analysis. Additionally, the spatial ranges of ancillary and fertility variables were calculated by a fitted semi-variogram model to provide a comparison of spatial distribution. Results indicated that ancillary imagery, captured in the low-frequency infrared spectrum, capture the spatial distribution of soil nutrients and may be used for predictive processes.

INTRODUCTION

Precision agriculture aims to increase the efficiency and yield of cropping systems by improving the spatial and temporal precision of inputs. To accurately estimate optimal seeding rates, variable rate irrigation, and variable soil amendment application it is critical that an accurate map of soil properties is attained. Often, maps of kriged soil properties are created by a variogram estimated from insufficient soil samples. This insufficiency could be caused by a small sample size or by sample locations that fail to adequately capture the spatial distribution of soil properties.

While these issues could be solved by dramatically increasing the sample count, the use of high-resolution and comprehensive soil sampling strategies is restrictive due to the prohibitive costs of sampling and analysis. To maintain a cost-effective strategy, a 2.5-acre grid has been accepted as the standard sampling scheme among farmers and consultants looking to implement precision agriculture technologies (Ferguson and Hergert, 2000; Godwin and Miller, 2003). While this strategy is certainly better than no sampling at all, it often fails to capture the true spatial distribution of soil properties. For precision agriculture purposes, it is essential that the sampling interval used captures the spatial dependence of soil characteristics, therefore allowing accurate kriging and mapping of fertility variables (Kerry and Oliver, 2008; Lopez-Granados et al., 2002). To avoid intensive sampling schemes, ancillary data must be used to supplement the 'gaps' in spatial data collected. The use of ancillary data allows for the detection of spatial patterns in soil fertility without the associated cost of fine-resolution sampling intervals. One readily accessible form of ancillary data that can be used to analyze soil properties is reflectance values recorded by spectroscopy.

The relationship between soil fertility and spectrometry has become an area of great interest in recent years. The ability to predict or analyze soil nutrient distribution by a non-invasive process has potential to drastically increase the efficiency and precision of soil amendment application. The progress of soil nutrient extraction has been well document since the first published acid-fluoride reagent extraction of soil P in the 1940's (Bray and Kurtz, 1945). Since this time the Mehlich 1 and 3 processes have become adopted as standards of nutrient extraction for their ability to give accurate results across a wide range of soil types (Mehlich, 1953; Mehlich, 1984). Unfortunately, it is only within recent decades that spectrometry technology has been developed to the point that multispectral sensors can accurately measure reflectance of color, near-infrared

(NIR), and short-wave-infrared (SWIR) wavelengths. The introduction of these technologies into soil fertility management has taken many forms. The in-lab spectral recording of soil samples has become a common practice and has been shown to detect soil texture, fertility, cation exchange capacity, and organic matter (Ben-Dor and Banin, 1995; Qiao and Zhang, 2012). Unfortunately, this method is limited in application as it requires the collection of soil samples and specialized laboratory equipment to record calibrated soil reflectance results.

An alternative method of spectral analysis can be conducted with multispectral imagery collected via aircraft and satellites. Remote sensing of this data allows for non-invasive large-scale analysis of soil properties. To determine the viability of remotely sensed multispectral data, spatial bi-correlation analysis will be used to detect those wavelengths which best capture the spatial variability present in soil fertility variables (K, P, and pH). This study seeks to analyze the usefulness of bi-correlation analysis of multispectral imagery (collected via satellite) and soil electrical conductivity to aid the creation of soil sampling schemes for precision agriculture purposes.

MATERIALS & METHODS

Description of Study Sites:

Two study sites were used for assessment of spatial variability of soil fertility using ancillary data classification. Both sites are currently used in field-crop production with corn being the primary product. The first site, hereby called ‘Samson’ is an 18.62-hectare field located at 31.115815, -86.093371, in Geneva County, Alabama near the city of Samson. The second site, hereby called ‘Autaugaville’ is a 41.56-hectare field located at 32.391148, -86.605381, in Autauga

County, Alabama near the town of Autaugaville. Table 2 records the soil series present at each site as well as the percentage of area. The study site locations are presented in Figure 1. Study site soil series maps are presented in Figure 2.

Soil Sampling

For both the 2018 and 2019 growing seasons, samples were collected in late winter from the Samson and Autaugaville study sites before soil amendments were applied. These samples were collected on a 50 m grid and were tested for P, K, magnesium, calcium, and zinc as well as soil pH, cation exchange capacity, and organic matter using the Mehlich I extraction method. Only soil P, K, and pH were used in this study. Soil P and K were selected due to their huge impact on crop growth, health, and yield (Prajapati, 2012; Ryan et al., 2012; Wang et al., 2013); and in consideration that they are the only two soil macronutrients to have standardized laboratory extraction methods and recommendations. Additionally, soil pH is studied due to the large influence it has over nutrient interactions, nutrient absorption, and plant health (Haynes and Naidu, 1998).

Samples were collected by bulk sampling within a 3 m radius of each grid center, with a collection depth of six inches. 156 samples were collected in Autaugaville in 2018 and 124 samples were collected in 2019. Sample size and study area was decreased in 2019 due to flood conditions in the Northeastern portion of the study site. The full sample size for 2018 data was used for spatial interpolation as this data contributed to the spatial dependence of soil variables. The Autaugaville area of interest (AOI) was restricted to the area sampled in the 2019 season to provide comparable spatial results between the two years. At the Samson study site, 101 samples were collected for both 2018 and 2019. Additionally, a 100 m grid was placed offset to the original 50 m grid to

increase the sample count and allow for more accurate capturing of varying spatial ranges at the Samson site. Maps of the sampling strategy at both Samson and Autaugaville are shown in Figure 3. Soil fertility distributions for Autaugaville are presented in Figure 5 and fertility distributions for Samson are presented in Figure 6.

Ancillary Data.

Remotely sensed imagery was sourced using USGS Earth Explorer. Satellite data was used from USGS' and NASA's Landsat Program (Landsat 7 and 8) as well as the ESA's Copernicus Program (Sentinel 2a). Theses satellites use multispectral sensors to record multiple bands of surface reflectance values which are corrected and calibrated by their respective space agency. Imagery was selected by two primary criteria. First, imagery captured near to the date of soil sampling was given preference. As the purpose of this study is to detect the significance of remotely sensed data on the creation of soil sampling schemes, using imagery that was taken soon before soil sampling is important as it will more accurately represent the current soil conditions that will be collected by soil sampling. Secondly, imagery was selected based on the lack of obstructions (cloud cover, vegetation, etc.) that may inhibit a clear recording of soil reflectance values. A table of Landsat 7, Landsat 8, and Sentinel 2a band wavelengths can be viewed in Table 3 and collection dates at each field can be viewed in Table 4. Figure 7 shows the relationship between bands among the three satellites.

In addition to remotely sensed data, spatial correlation results were also analyzed between soil sampling results and another commonly recorded parameters, Soil EC. Soil EC was collected at both Samson and Autaugaville in 2016 using a Veris 3100 equipped with a real-time kinematic

GPS to record Shallow EC (approximately 1 ft.), Deep EC (approximately 3 ft), and elevation values. Soil EC data is mapped for Autaugaville and Samson in Figure 4.

Data Preprocessing:

Before analysis, all geospatial data were converted to the Universal Transverse Mercator (UTM) coordinate system in the North American Datum of 1983 (NAD83), which uses meters as the standard unit of measurement. As both sites are located in Alabama, UTM zone 16N was used. Study sites are relatively close to the standard meridian of the projection zone; therefore, the spatial data is assumed to contain minimal scale distortion and can be accurately used for spatial analysis. Geographic transformations, tabular analysis, and spatial data creation were completed using ESRI's ArcGIS 10.4.1 for Desktop. Data clipping and organization were completed using GDAL 2.3.2 from the Open Source Geospatial Foundation.

Spatial Interpolation:

As a result of the various collection resolutions used to create data it was necessary that all variables were transformed to the same spatial scale before spatial correlation analysis. To accomplish this, an overlay grid was created for each study site. This polygon grid is necessary for the interpolation of original data to a standard scale. Due to the possibility of over-generalizing, a grid size must be defined that is fine enough to not eliminate the small-scale spatial variability present in original data. A polygon size of 10 m x 10 m was determined to accurately portray the spatial variability present while also reducing processing time and prediction errors from interpolation. The 10 m grids for both Autaugaville and Samson sites can be seen in Figure 3. Values for P, K, soil pH, soil EC (shallow and deep), Landsat 7 bands 1 – 7, Landsat 8 bands 1 –

7, and Sentinel 2a bands 1 – 12 were estimated for each grid polygon by ordinary kriging interpolation. Ordinary kriging was performed in SpaceStat 4.0.21 by BioMedware Inc. (https://www.biomedware.com/software/spacestat/). This interpolation process effectively resamples each variable into the 10 m spatial scale, resulting in a standard sample size for all variables. The polygon count is 4,387 for Autaugaville and 2,000 for Samson.

Spatial interpolation by kriging is done by assigning weights to observed values based on their distance from the point that is to be interpolated. The process of kriging estimates the value of an unobserved location by interpolating from weighted points within the determined range of the semi-variogram model. Ordinary kriging differs from other spatial interpolation methods (such as Inverse Distance Weighted, IDW) in that three major parameters are used to guide the kriging process. The range, or the distance at which auto-correlation between points can no longer be observed, is used to determine the suggested search distance for the interpolation process. The nugget is a measure of variation in observations that occurs at distances smaller than the sampling interval. Lastly, the sill represents the variation in observations that occurs at the range, where autocorrelation is no longer observed.

Ordinary kriging differs from other spatial interpolation methods (such as Inverse Distance Weighted, IDW) in that it will only interpolate new data using observations within the threshold distance of spatial autocorrelation, called the range. The range of each variable is determined by calculating the empirical semi-variogram of the data, which plots the Euclidean distances of observed data pairs against the recorded difference between data pairs. The semi-variogram $\gamma(h)$, which is defined by Matheron (1963) as half the average squared difference between points separated at distance h. The empirical semi-variogram is calculated as:

$$\gamma(h) = \frac{1}{2|N(h)|}\sum_{N(h)}(z_i - z_j)^2$$

where *N(h)* is the set of all pairwise Euclidean distances $i - j = h$, $|N(h)|$ is the number of distinct pairs in *N(h)*, and z_i *and* z_j are data values at spatial location *i* and *j*, respectively. *h* represents a distance measure with magnitude only (Ford, 2010).

Typically, semi-variance observations are grouped into bins by lag size selection, $h \pm \delta$ rather than using exact distances. Therefore, the empirical semi-variograms $\hat{\gamma}(h \pm \delta)$ can be calculated for each bin:

$$\hat{\gamma}(h \pm \delta) = \frac{1}{2|N(h \pm \delta)|}\sum_{(i,j)\epsilon\, N(h \pm \delta)}|z_i - z_j|^2$$

Each pair of points separated by *h* plus or minus lag size range δ are found. These form the set of points $N(h \pm \delta) \equiv \{(s_i, s_j) : |s_i, s_j| = h \pm \delta; i, j = 1, \dots, N\}$. The number of points in this bin is $N(h \pm \delta)$. For each pair of points *i, j*, the square of the difference in the observation is found $(|z_i - z_j|)^2$. These squared differences are added together and normalized by the natural number$|N(h \pm \delta)|$. The result is then divided by 2 for the semi-variogram.

A model is then fit to best capture the spatially dependent variation of the dataset by minimizing residuals and predictive error. Often, the fit of the model can be improved by the addition of a second model, allowing for each model to better fit part of the data (short and long spatial ranges), this linear combination of models is called a nested model. Not every semi-variogram is best captured by a nest model and may instead be best represented by a single model. Two important parameters for the semi-variogram are the nugget and sill. The sill is defined as the difference in observation values at the range. The nugget effect is the phenomena which results in

a difference value greater than zero, even at an infinitesimally small separation distance. The nugget to sill ratio is often used to indicate what fraction of overall variance is present at a distance smaller than the smallest separation distance. This ratio indicates how much variance has been accounted for by the model (Allington, 2010).

The nugget to sill ratio and range created from the semi-variogram are recorded from Autaugaville data in 2018 (Table 11) and 2019 (Table 12) and Samson data in 2018 (Table 13) and 2019 (Table 14).

Correlation Analysis:

The Pearson product-moment correlation coefficient was calculated for both years at each study site. This relationship represents the strength of a linear relationship between soil tested variables and ancillary data. For this study K, P, soil pH, and soil EC were used as dependent variables. The strength of correlation was measured between these dependent variables and the ancillary data: Landsat 7 bands 1 - 7, Landsat 8 bands 1 - 7, Sentinel bands 1 - 12, and soil EC. Correlation relationships were calculated using Python 2.7 and the Pandas package.

RESULTS AND DISCUSSION

Soil Attributes Summary Statistics:

The summary statistics of the soil P, K, and soil pH measured in 2018 and 2019 at both study sites are presented in Table 5. Autaugaville displayed relatively inconsistent variation in soil variables between 2018 and 2019. Soil K σ changed by 11.76 lb./acre, soil P σ changed by 15.72 lb./acre, and soil pH σ changed by 0.08. For every variable, σ was higher for 2019 than the previous year. This increase in variation could be attributed to heavier rainfall during the 2018 growing

season, likely increasing the runoff and seepage of applied soil nutrients in portions of the field more suited for collecting water.

By comparison, Samson soil variables showed very little variation between the 2018 and 2019 testing dates. Soil K σ only changed by 4.01 lb./acre, soil P σ changed by just 1.31 lb./acre, and soil pH σ changed by 0.03. This, in combination with the relatively stable Min – Max values, indicates that soil conditions and treatments were very similar for both years.

In general, the distributions are symmetric with only one variable, Autaugaville P, showing a skewness exceeding 1.0 in both 2018 and 2019. Despite the slight skewness observed, these observations were not transformed as the use of a logarithm did little to change the skewness of the distributions. The distribution of Samson K in 2019 appears to be somewhat bimodal, though all others display a unimodal distribution.

The product moment correlation coefficients between soil variables, by year, all displayed a weak trend, as seen in the matrix at Table 6. At the Autaugaville site, all of the significant correlations, though relatively weak, were positive. This positive relationship results in the presence of "pockets" of high or low soil fertility. It is possible that this effect is caused by an external limiting factor (I.e. water availability or over-saturation at this non-irrigated site) which prohibited the full uptake of available soil nutrients by crops. Conversely, at the Samson site, P and K displayed a significant negative relationship during both years, although these correlations were relatively weak.

Correlation and Semi-variogram Analysis:

Correlation results for soil fertility and satellite spectral bands are presented in Tables 7 (Autaugaville, 2018), 8 (Autaugaville 2019), 9 (Samson 2018), and 10 (Samson 2019). In general,

the strongest relationships are seen between soil fertility variables and spectral bands in the Near-Infrared (NIR), Shortwave-Infrared (SWIR), and Vegetation-Red-Edge (VRE) spectrums. These three portions of the electromagnetic spectrum each contain a different section of low-frequency infrared reflectance. It is also expected that if two variables are correlated they will also exhibit a similar spatial variability and range. A comparison of the empirical and fitted experimental semi-variograms allows for analysis of spatial behavior. Fitted semi-variogram parameters are provided in Table 11 (Autaugaville 2018), 12 (Autaugaville 2019), 13 (Samson 2018) and 14 (Samson 2019). Results are presented here by ancillary data type, further divided by study site and sampling year.

Landsat 7 and Soil K:

Correlations between Landsat 7 reflectance values and soil macronutrients was completed for the Autaugaville study site in both 2018 and 2019, and the Samson site in 2019. At the Autaugaville site, bands 1 (blue) and 3 (red) displayed moderate correlations to soil K in both 2018 and 2019. Additionally, band 4 (NIR) displayed a strong correlation to soil K in 2018 and a moderate correlation in 2019. This correlation with reflectance from the NIR range is consistent with the reflectance of potassium in both soils and crop canopy (Yu et al., 2016; Pimstein et al., 2011).

At the Samson site, soil K correlations were consistently higher than those recorded at Autaugaville, likely due to the better conditions under which imagery was collected. Strong relationships were observed for bands 1 (blue), 3 (red), 5 (SWIR1), and 7 (SWIR2). The strengthened relationship in the SWIR range (1500 – 2350 nm) is also consistent with the results published by Yu et al. (2016) who found a strong relationship between soil K and indices using

SWIR. The ranges of the red and SWIR1 bands (258 m and 299 m), by fitted variogram model, are also relatively similar to the range fitted for soil K (243 m) indicating that they share a common spatial variability with K. The consistency of relationship strength across the two years, and two study sites, indicates that these bands (blue, red, NIR, SWIR) have potential to be used for the remote detection of soil K. It is likely that the creation of a potassium-specific indices, using a linear combination of red, blue, and NIR, could result in a stronger relationship and greater predictive power.

Landsat 7 and Soil P:

Similar to soil K, soil P was found to have moderate relationships with Landsat 7 bands 1 (blue), 2 (green), 3 (red), and 7 (SWIR1) at the Autaugaville study site. Additionally, the range for 2018 soil P (2411 m) and band 7 (2344 m) were similar as were 2019 soil P (78 m) and band 4 (98 m). Unfortunately, none of these moderate relationships were consistent between the two years sampled. When it comes to predictive power, this inconsistency is just as damaging as weak correlative relationships. A lack of temporal steadiness points towards inconsequential correlations and non-repeatable results. On the other hand, relationships from 2019 imagery at the Samson study site were strong for every band besides band 4 (NIR). It is suspected that, once again, the increased correlation is a result of better field conditions during the capturing of Samson imagery, allowing for a better recording of bare-soil reflectance. This indicates that soil P, or an external variable controlling the distribution of soil P, is not detectable by Landsat 7 sensors when crop residue is present.

Landsat 7 and soil pH:

Landsat 7 data failed to produce any moderate strength correlation coefficients with Autaugaville soil pH for both 2018 and 2019, and with Samson soil in 2019. This comprehensive lack of correlative strength implies that Landsat 7 sensors are not appropriate for the detection of soil variables responsible for pH. This is not to say that the spatial distribution of soil pH is undetectable by remotely sensed imagery; rather that the band widths, method of capture, or preprocessing of Landsat 7 imagery inhibit this detection.

Landsat 8 and Soil K:

Correlation coefficients returned for soil K in 2018 at the Autaugaville study site contained no moderate or strong relationships. While it cannot be expressed with certainty, it is likely that the field conditions at the time of imagery capture prohibited a quality recording of soil reflectance. This is further strengthened by the lack of relationship between Landsat 8 data and soil P and soil pH for Autaugaville 2018. This lack of relationship, which differs from the results attained from 2019 and the Samson study site indicate that noise was introduced into the imagery masking its relationship with soil variables. In contrast to this, Landsat 8 imagery from 2019 showed moderate relationships to soil K in bands 5 (NIR), 6 (SWIR1), and 7 (SWIR2). Unfortunately, none of these bands also exhibited similarity to the spatial range from the fitted variogram model for soil K.

Soil K data from the Samson study site in 2018 displayed moderate negative relationships with Landsat 8 bands 2 (blue), 3 (green), and 4 (red) and additional strong negative relationships with bands 6 (SWIR1) and 7 (SWIR2). These are similar to the results calculated from soil K data in the following year, 2019, which resulted in the strongest relationships with bands 2 (blue), 6 (SWIR1), and 7 (SWIR2). It was also observed that the 2018 soil K spatial range (74 m) was similar to the range of band 3 (86 m) and 7 (89 m). Spatial ranges similar to 2019 soil K (244 m)

included bands 1 (234 m) and 7 (237 m) in 2019. The strong correlation and similar spatial ranges indicate that these reflectance values will be useful for the prediction of soil K spatial distribution.

Landsat 8 and Soil P:

Similar to the results reported for soil K, Landsat 8 data correlated poorly with soil P from the Autaugaville study site in 2018 and resulted in moderate positive correlations for bands 5 (NIR), 6 (SWIR1), and 7 (SWIR2) in 2019. It was observed that the spatial range for soil P in 2019 (79 m) and band 4 (98 m) were very similar

At the Samson study site, in 2018, Landsat 8 bands 6 (SWIR1) and 7 (SWIR2) displayed moderate positive relationships with soil P. In 2019, the strongest positive relationships were once again produced with bands 6 (SWIR1) and 7 (SWIR2). In 2018, the range of soil P (86 m) was very close to that shown by band 7 (89 m), yet in 2019 the range of soil P was observed to have changed to 145 m, thus making it closer to the 112 m range observed for band 6. It is likely that both of the SWIR bands from Landsat 8 can be used for the mapping of soil P distributions.

Landsat 8 and Soil pH:

Interestingly, Landsat 8 data correlated only weakly with soil pH values at both study sites for both years, except for Autaugaville 2018, were every band displayed a moderate relationship. It is suspected that this is also caused by the suboptimal recording of reflectance values at the time of imagery capture. These faulty recordings highlight the importance of clear and well-timed imagery for soil fertility mapping.

Sentinel 2a and Soil K:

At the Autaugaville study site, in 2018, soil K displayed a moderate negative relationship with Sentinel band 2 (blue) and a moderate positive relationship with bands 6 (VRE2), 7 (VRE3), and 8 (NIR). Despite their correlations, these bands did not show similarity to the spatial range of soil K. For the 2019 soil K samples, only Sentinel band 2 (blue) returned a moderate negative correlation. This band also failed to mimic the spatial range established by soil K distribution.

At the Samson study site, in 2018, soil K displayed a strong negative relationship with bands 11 (SWIR1) and 12 (SWIR2). This relationship is confirmed by the similar spatial range of band 12 (76 m) to soil K (74 m). Similar results are observed in 2019 with strong negative correlations for bands 11 (SWIR1) and 12 (SWIR2) followed by bands 2 (blue) and 3 (green). These repeated relationships and similarities in range suggest that the spectrum captured by Sentinel bands 11 and 12, supplemented by blue reflectance, and could be used for the prediction of soil K variability.

Sentinel 2a and Soil P:

No strong relationships were observed between data collected by Sentinel sensors and Autaugaville soil P. A moderate positive relationship was detected from band 11 (SWIR1) in 2018 and band 2 (blue) in 2019. Due to the inconsistency across years and the lack of similarity in spatial ranges, it is unlikely that the Sentinel imagery captured the spatial distribution of soil P.

At the Samson study site, moderate positive relationships were displayed by bands 11 (SWIR1) and 12 (SWIR2) in 2018, which also displayed strong positive relationships in 2019. Additionally, the range calculated for soil P in 2018 was 86 m, and the range of band 12 that same year was 76 m. This indicates that the same bands which likely capture the distribution of soil K also capture soil P, though the relationships are inversed.

Sentinel 2a and Soil pH:

All Sentinel bands failed to result in anything greater than a weak relationship with the exception of bands 2 (blue) and 3 (green) at the Samson study site in 2018. Due to the general lack of correlation across both study sites and both years, it is impossible to confidently declare any remote data from Sentinel 2a fit for soil pH distribution prediction.

Soil EC and Soil K:

At the Autaugaville study site in 2018, Shallow soil EC displayed a moderate negative relationship to soil K. In 2019, Shallow soil EC once again outperformed Deep soil EC and returned a strong negative correlation. Interestingly, these results were reversed for the Samson study site, where both shallow and deep soil EC displayed strong positive correlations. Heiniger et al. (2003) reported that, "If we assume that salinity is not a major factor in most productive agricultural fields, it follows that in nonsaturated conditions, changes in soil nutrient levels will most likely influence EC_a by changing ECs through differences in the type and number of cations held by the soil particles." It is therefore possible that the inverse relationships observed at the two study sites is caused by a difference in field conditions at the time of EC sampling.

As previously noted, the Autaugaville study site is located along a river and experiences frequent flooding conditions. Approximately 44% of the field is composed of McQueen series soil, which is described as slow permeability with a water table perched at a depth of 4 to 6 feet during winter and spring of most years (NRCS, 1998). An additional 32% of the field contains soil from the Roanoke series which is classified as "poorly drained" with slow to very slow permeability (NRCS, 2006). It is possible that saturated field conditions have impaired the detection of soil K

by reporting highly conductive results in areas with low K. This theory is supported by the relative lack of similarity between soil K and soil EC ranges at the Autaugaville site. On the other hand, the ranges for soil K (244 m) and shallow soil EC (258 m) were very similar for Samson 2019. This, along with a strong correlative relationship, suggests that soil EC may be used as an effective variable for the detection of soil K variability in non-saturated fields.

Soil EC and Soil P:

Shallow soil EC returned only moderate negative relationships with soil P for both years at the Autaugaville site. These variables were not found to display a similar spatial range to that calculated for soil P in either year. At the Samson study site, strong correlations were detected for both shallow and deep soil EC in both years, with shallow soil EC displaying a slightly stronger relationship. The spatial ranges for deep soil EC (189 m) and soil P (145 m) were relatively similar and imply potential predictive power.

Soil EC and Soil pH:

Both shallow and deep soil EC returned only weak negative correlation with soil pH at the Autaugaville study site for both 2018 and 2019. These results were very similar to those attained by soil K and likely suffer from the same issue of saturation masking. Also similar to soil K was the strong positive relationships observed for deep and shallow soil EC in 2018 at Samson followed by moderate positive relationships in 2019. The spatial range of deep soil EC (626 m) and soil pH (563 m) may imply a vague similarity of spatial distribution but fails to identify a primary candidate for soil pH distribution prediction.

Ancillary Data Combinations:

By using spatial bi-correlation analysis and semi-variogram analysis to determine those ancillary data which accurately display the distribution of soil fertility variables, combinations can be created to further increase the accuracy and certainty of distribution prediction. The condensing of overlapping bands helps to clarify the specific wavelengths that are shown to most accurately portray soil variables. For K, the wavelengths 730 – 900 nm, 1540 – 1680 nm, and 2080 – 2300 nm tend to best relate soil spatial distribution. Soil P is better represented by 450 – 590 nm, 1540 – 1680 nm, and 2080 – 2300 nm. Lastly, soil pH is distributed similarly to 450 – 580 nm and 740 – 880 nm.

CONCLUSION

Correlation indicates that remotely sensed multispectral imagery can serve as ancillary data for the prediction of soil fertility spatial ranges. While moderate relationships were detected across the board, especially among wavelengths in the near-infrared and shortwave-infrared ranges, it is difficult to find consistent correlation patterns across the two study sites and two years included in this study. In general, the Samson site displayed stronger relationships, especially for soil P and soil K. The cropping system used at the Samson site involves the use of traditional tillage and early spraying of winter cover crops. This practice allows for a greater window in which satellite can be captured of bare soil reflectance. It is likely that the decreased strength of relationships detected at the Autaugaville site are the result of erroneous reflectance values caused by crop residue, natural field weed conditions, and lack of soil disturbance at the time of imagery capture. This change in relationship strength indicates that the use of multispectral imagery at any given field could be limited by the tillage and cover practices implemented. Fortunately, soil EC detection is not dependent on cover practices. Additionally, the semi-permanence of soil EC, as a pseudo-sampling

of soil texture values, makes it a more universal resource for the detection of soil fertility spatial ranges.

It can be concluded that the combination of remotely sensed multispectral imagery and soil EC may serve as an effective form of ancillary data for the creation of soil sampling strategies. This can be accomplished by the creation of soil nutrient predications, an analysis of prediction accuracy, and the resulting minimum number of samples which could be collected to improve the accuracy of spatial predictions. For future research, it is recommended that these data sources continue to be analyzed as a set of independent data for soil nutrient prediction. Possible methodologies include spatial regression, cokriging, or unsupervised classification to create accurate nutrient prescription maps from relatively sparse sampling data. In the next chapter, we will explore the use of principal component analysis as a method of data dimension reduction to allow for the use of K-means classification as a prediction of soil fertility zones.

TABLES

Table 2. Study Site Soil Series

Samson Soils

Soil Series	% of AOI	Taxonomic Class
Alpin Sand	20.6	Thermic, coated Lamellic Quartzipsamments
Eunola Sandy Loam	69.7	Fine-loamy, siliceous, semiactive, thermic Aquic Hapludults
Leaf-Lenoir Complex	9.7	Fine, mixed, semiactive, theric Aeric Paleaquults

Autaugaville Soils

Soil Series	% of AOI	Taxonomic Class
Altavista Loam	9.9	Fine-loamy, mixed, semiactive, thermic Aquic Hapludults
Blanton Loamy Sand	1.7	Loamy, siliceous, semiactive, thermic Grossarenic Paleudults
Lakeland Loamy Sand	10.3	Thermic, coated Typic Quartzipsamments
McQueen Silt Loam	43.7	Fine, mixed, semiactive, thermic Typic Hapludults
Roanoke Complex	32.0	Fine, mixed, semiactive, thermic Typic Endoaquults
Wickham Fine Sandy Loam	2.3	Fine-loamy, mixed, semiactive, thermic Typic Hapludults

Table 3. Description of Remote Data Wavelengths

Satellite	Band #	Description	Spectral range (nm)
Landsat 7	Band 01	Blue	450-520
	Band 02	Green	520-600
	Band 03	Red	630-690
	Band 04	Near Infrared	770–900
	Band 05	Shortwave Infrared 1	1550-1750
	Band 06	Thermal Infrared	10400-12500
	Band 07	Shortwave Infrared 2	2090-2350
Landsat 8	Band 01	Ultra-Blue	435-451
	Band 02	Blue	452–512
	Band 03	Green	533–590
	Band 04	Red	636-673
	Band 05	Near Infrared	851–879
	Band 06	Shortwave Infrared 1	1566–1651
	Band 07	Shortwave Infrared 2	2107–2294
Sentinel 2A	Band 01	Coastal Aerosol	430.4-45734
	Band 02	Blue	447.6–545.6
	Band 03	Green	537.5–582.5
	Band 04	Red	647.5-681.5
	Band 05	Vegetation Red Edge	694.4-713.4
	Band 06	Vegetation Red Edge	731.2–749.2
	Band 07	Vegetation Red Edge	768.5–796.5
	Band 08	Near Infrared	762.6–907.6
	Band 09	Narrow Near Infrared	932-958
	Band 10	Shortwave Infrared Cirrus	1336-1411
	Band 11	Shortwave Infrared 1	1542.2–1685.2
	Band 12	Shortwave Infrared 2	2081.4–2323.4

Table 4. Remote Sensing Data Collection Dates

Site	Year	Satellite	Collection Date
Autaugaville	*2018*	Landsat 7	Monday, February 19, 2018
		Landsat 8	Wednesday, March 7, 2018
		Sentinel 2a	Saturday, March 31, 2018
	2019	Landsat 7	Friday, January 11, 2019
		Landsat 8	Tuesday, December 18, 2018
		Sentinel 2a	Friday, March 29, 2019
Samson	*2018*	Landsat 7	Wednesday, January 17, 2018
		Landsat 8	Thursday, January 25, 2018
		Sentinel 2a	Monday, February 19, 2018
	2019	Landsat 7	Friday, January 11, 2019
		Landsat 8	Monday, January 28, 2019
		Sentinel 2a	Thursday, March 21, 2019

Table 5. Descriptive statistics of soil fertility samples collected in 2018 and 2019

All samples composited by bulk sampling at a 6" depth and extracted with Mehlich III extractant. All results are given in lb/acre.

Variable		Descriptive Statistics (K/P results given in lb./acre)					
		Mean	Min - Max	S.D	Variance	Kurtosis	Skewness
Autaugaville							
2018							
	K	152.14	58 - 293	41.70	1739.00	0.74	0.63
	P	20.93	4 - 135	9.37	87.80	24.45	1.13
	pH	5.95	4.7 - 7.7	0.43	0.19	1.50	0.42
2019							
	K	163.05	62 - 408	53.46	2858.00	2.08	0.65
	P	34.71	6 - 124	25.09	629.46	3.09	1.86
	pH	6.47	4.8 - 7.9	0.51	0.26	0.43	-0.11
Samson							
2018							
	K	179.50	47 - 370	65.27	4260.00	-0.26	0.23
	P	61.65	9 - 130	31.21	974.17	-1.09	-0.06
	pH	6.23	5.2 - 7.6	0.45	0.20	0.27	0.24
2019							
	K	181.48	52 - 357	69.28	4800.00	-0.68	0.09
	P	65.83	13 - 144	32.52	1057.00	-0.58	0.46
	pH	5.83	5.1 - 7.6	0.42	0.17	2.06	0.90

Table 6. Soil fertility samples correlation matrices

Autaugaville 2018		
	K	**P**
pH	-0.05	0.11

Autaugaville 2019		
	K	**P**
pH	0.29*	0.49**

Samson 2018		
	K	**P**
pH	0.30*	-0.05

Samson 2019		
	K	**P**
pH	0.33*	0.33

*, significant for $P < 0.01$; ** significant for $P < 0.0001$, all other correlation coefficients are not significant.

Table 7. Correlations between soil fertility and ancillary variables. Autaugaville 2018

Ancillary Data	Band	K	P	pH	Description
Landsat 7	1	**-0.317****	-0.126**	0.084**	Blue
	2	-0.258**	0.447**	-0.005	Green
	3	**-0.317****	**0.366****	0.022	Red
	4	**0.538****	-0.160**	-0.194**	Near Infrared
	5	0.060**	-0.118**	-0.021	Shortwave Infrared 1
	7	-0.029	**0.411****	0.039*	Shortwave Infrared 2
Landsat 8	2	-0.097**	0.024	**0.346****	Blue
	3	-0.092**	0.045*	**0.331****	Green
	4	-0.101**	0.048*	**0.342****	Red
	5	-0.120**	0.081**	**0.381****	Near Infrared
	6	-0.127**	0.098**	**0.368****	Shortwave Infrared 1
	7	-0.150**	0.048*	**0.359****	Shortwave Infrared 2
Sentinel 2a	2	-0.277**	-0.249**	0.061**	Blue
	3	-0.038	-0.106**	0.047*	Green
	4	-0.199**	-0.225**	-0.007	Red
	5	0.125**	0.016	0.058*	Vegetation Red Edge
	6	0.385**	0.155**	0.046*	Vegetation Red Edge
	7	**0.409****	0.181**	0.027*	Vegetation Red Edge
	8	**0.349****	0.182**	-0.004	Near Infrared
	11	0.119**	**0.309****	0.020	Shortwave Infrared 1
	12	0.018	0.286**	0.072**	Shortwave Infrared 2
EC	Deep	-0.126**	-0.202**	-0.083**	EC Deep
	Shallow	**-0.319****	**-0.471****	-0.069**	EC Shallow

*, significant for P < 0.01; ** significant for P < 0.0001, all other correlation coefficients are not significant. Moderate-to-Strong relationship in **Bold**.

Table 8. Correlations between soil fertility and ancillary variables. Autaugaville 2019

Ancillary Data	Band	K	P	pH	Description
Landsat 7	1	**-0.338****	**-0.375****	-0.124**	Blue
	2	-0.215**	-0.265**	-0.120**	Green
	3	**-0.336****	-0.263**	-0.101**	Red
	4	**0.430****	0.148**	0.086**	Near Infrared
	5	0.057**	0.062**	0.062**	Shortwave Infrared 1
	7	-0.154**	-0.018	0.058*	Shortwave Infrared 2
Landsat 8	2	0.188**	0.181**	0.136**	Blue
	3	0.139**	0.035	0.022	Green
	4	-0.035	-0.053*	-0.012	Red
	5	**0.472****	0.278**	0.176**	Near Infrared
	6	**0.441****	**0.381****	0.244**	Shortwave Infrared 1
	7	**0.408****	**0.378****	0.252**	Shortwave Infrared 2
Sentinel 2a	2	**-0.403****	**-0.352****	-0.126**	Blue
	3	-0.250**	-0.244**	-0.101**	Green
	4	-0.055**	-0.150**	-0.139**	Red
	5	-0.099**	-0.142**	-0.039*	Vegetation Red Edge
	6	-0.034	0.035	0.162**	Vegetation Red Edge
	7	0.027	0.083**	0.190**	Vegetation Red Edge
	8	0.023	0.084**	0.164**	Near Infrared
	11	0.151**	0.093**	-0.083**	Shortwave Infrared 1
	12	0.117**	0.084**	-0.086**	Shortwave Infrared 2
EC	Deep	**-0.337****	-0.195**	-0.062**	EC Deep
	Shallow	**-0.567****	**-0.420****	-0.201**	EC Shallow

*, significant for $P < 0.01$; ** significant for $P < 0.0001$, all other correlation coefficients are not significant. Moderate-to-Strong relationship in **Bold**.

Table 9. Correlations between soil fertility and ancillary variables. Samson 2018

Ancillary Data	Band	K	P	pH	Description
Landsat 8	2	**-0.493****	0.219**	-0.008	Blue
	3	**-0.405****	0.145**	0.081	Green
	4	**-0.423****	0.201**	0.127**	Red
	5	-0.017	-0.144**	-0.163**	Near Infrared
	6	**-0.644****	**0.394****	-0.154**	Shortwave Infrared 1
	7	**-0.633****	**0.392****	-0.093**	Shortwave Infrared 2
Sentinel 2a	2	0.114**	-0.143**	**0.369****	Blue
	3	0.004**	-0.091**	**0.312****	Green
	4	-0.182**	0.051**	0.202**	Red
	5	-0.192**	0.037**	0.184**	Vegetation Red Edge
	6	0.091**	-0.180**	0.295**	Vegetation Red Edge
	7	0.126**	-0.210**	0.289**	Vegetation Red Edge
	8	0.073**	-0.158**	0.251**	Near Infrared
	11	**-0.519****	**0.346****	-0.069**	Shortwave Infrared 1
	12	**-0.548****	**0.387****	-0.095**	Shortwave Infrared 2
EC	Deep	**0.680****	**-0.561****	**0.528****	EC Deep
	Shallow	**0.675****	**-0.579****	**0.538****	EC Shallow

*, significant for $P < 0.01$; ** significant for $P < 0.0001$, all other correlation coefficients are not significant. Moderate-to-Strong relationship in **Bold**.

Table 10. Correlations between soil fertility and ancillary variables. Samson 2019

Ancillary Data	Band	K	P	pH	Description
Landsat 7	1	**-0.580****	**0.533****	-0.083*	Blue
	2	**-0.478****	**0.507****	0.003	Green
	3	**-0.522****	**0.527****	-0.001	Red
	4	0.180	-0.042**	0.103**	Near Infrared
	5	**-0.589****	**0.565****	-0.101**	Shortwave Infrared 1
	7	**-0.591****	**0.576****	-0.058*	Shortwave Infrared 2
Landsat 8	2	**-0.559****	**0.602****	0.061*	Blue
	3	**-0.530****	**0.553****	0.059*	Green
	4	**-0.581****	**0.574****	0.041	Red
	5	0.221**	-0.148**	0.026	Near Infrared
	6	**-0.686****	**0.626****	-0.102**	Shortwave Infrared 1
	7	**-0.678****	**0.625****	-0.072*	Shortwave Infrared 2
Sentinel 2a	2	**-0.616****	**0.497****	-0.257**	Blue
	3	**-0.517****	**0.439****	-0.185**	Green
	4	-0.349**	0.294**	-0.153**	Red
	5	-0.268**	0.242**	-0.135**	Vegetation Red Edge
	6	-0.277**	0.264**	-0.144**	Vegetation Red Edge
	7	-0.254**	0.244**	-0.163**	Vegetation Red Edge
	8	-0.261**	0.245**	-0.176**	Near Infrared
	11	**-0.644****	**0.546****	-0.267**	Shortwave Infrared 1
	12	**-0.685****	**0.588****	-0.235**	Shortwave Infrared 2
EC	Deep	**0.751****	**-0.507****	**0.480****	EC Deep
	Shallow	**0.725****	**-0.568****	**0.458****	EC Shallow

*, significant for $P < 0.01$; ** significant for $P < 0.0001$, all other correlation coefficients are not significant. Moderate-to-Strong relationship in **Bold**.

Table 11. Parameters of fitted variogram models for Autaugaville 2018

Variable	Nugget	Sill 1	Sill 2	Range 1	Range 2	Nugget:Sill	MSS
P	140.29	33.46	957.01	60	2411	12%	0.01
K	1082.79	760.39		234		59%	0.02
Soil pH	0.17	0.02	0.04	852	853	73%	0.01
EC_Shallow	4.93	5.16		279		49%	0.01
EC_Deep	1.32	42.81	23.06	72	1252	2%	0.00
L7_B01	770.15	816.30	2019.96	133	456	21%	0.02
L7_B02	1357.87	1154.02	2160.19	168	482	29%	0.02
L7_B03	2875.21	9843.82		386		23%	0.01
L7_B04	2838.97	15599.17	26680.42	86	814	6%	0.01
L7_B05	67.15	23775.71	186491.40	102	2361	0%	0.01
L7_B07	9748.36	3265.90	97966.61	633	2344	9%	0.01
L8_B02	782.02	1154632.10	257063.17	356	456	0%	0.04
L8_B03	0.12	1201568.94	321555.84	352	458	0%	0.04
L8_B04	0.02	1364325.20	328085.68	349	462	0%	0.04
L8_B05	2.00	1308488.40	250940.49	340	446	0%	0.06
L8_B06	0.08	1585038.06		333		0%	0.06
L8_B07	1.08	1534677.52		332		0%	0.06
S2_B02	1701.83	860.12	1140.33	184	499	46%	0.00
S2_B03	1844.16	2943.01	532.38	178	586	35%	0.00
S2_B04	8112.98	3476.82	5454.68	197	498	48%	0.00
S2_B05	4071.34	9683.14	5176.74	204	2795	22%	0.00
S2_B06	8485.11	10323.09	60014.58	161	2762	11%	0.00
S2_B07	12414.76	10716.53	86043.96	134	2762	11%	0.00
S2_B08	23350.75	15149.98	39616.79	147	2836	30%	0.01
S2_B11	8865.05	29095.90	94286.77	130	2803	7%	0.00
S2_B12	7828.41	13638.03	126225.97	100	2579	5%	0.00

Ranges given in meters.

Table 12. Parameters of fitted variogram models for Autaugaville 2019

Variable	Nugget	Sill 1	Sill 2	Range 1	Range 2	Nugget:Sill	MSS
P	257.61	185.78	2570.83	79	739	9%	0.01
K	2160.57	1598.58		240		57%	0.01
Soil pH	0.00	0.19	0.68	77	1019	0%	0.01
EC_Shallow	4.64	3.36	2.13	146	470	46%	
EC_Deep	30.13	15.25	47.34	149	1827	32%	
L7_B01	746.63	2503.60		170		23%	0.01
L7_B02	279.37	5218.54	18280.90	176	2691	1%	0.03
L7_B03	121.50	12121.66		170		1%	0.01
L7_B04	10213.69	28159.77	134905.20	98	778	6%	0.01
L7_B05	17379.49	27223.63		235		39%	0.01
L7_B07	82.22	23707.62		146		0%	0.02
L8_B02	665.22	1279.87	6667.79	138	2809	8%	0.01
L8_B03	954.52	3603.81	32799.89	118	2407	3%	0.01
L8_B04	1085.01	10533.39	45168.01	126	2777	2%	0.01
L8_B05	0.35	48280.47	74474.19	84	792	0%	0.01
L8_B06	12.19	62010.57	25266.10	113	536	0%	0.01
L8_B07	3410.05	26097.92	74107.59	148	2808	3%	0.01
S2_B02	848.98	2210.07	595.47	156	2318	23%	0.02
S2_B03	916.04	4001.50	699.48	176	2009	16%	0.03
S2_B04	308.98	9938.48	27213.81	239	2854	1%	0.00
S2_B05	1255.73	10517.82		204		11%	0.01
S2_B06	3952.64	11806.42		170		25%	0.08
S2_B07	3122.11	16567.17		148		16%	0.09
S2_B08	6901.75	16560.78		149		29%	0.10
S2_B11	2732.91	46611.39		436		6%	0.00
S2_B12	659.24	3028.65	22980.63	86	468	2%	0.00

Ranges given in meters.

Table 13. Parameters of fitted variogram models for Samson 2018

Source	Nugget	Sill 1	Sill 2	Range 1	Range 2	Nugget:Sill	MSS
P	161.41	669.12	483.96	89	544	12%	0.00
K	0.02	2357.39	2969.97	60	330	0%	0.01
Soil pH	0.10	0.04	0.08	30	425	44%	0.00
EC_Shallow	13.24	5.54	102.53	108	2067	11%	0.00
EC_Deep	37.46	52.57	482.84	93	2115	7%	0.02
L8_B02	755.67	755.67	37856.65	110	284	2%	0.01
L8_B03	6306.11	54821.72	50474.58	113	300	6%	0.00
L8_B04	9561.67	97385.20	190425.85	116	268	3%	0.01
L8_B05	0.13	372216.24		103		0%	0.02
L8_B06	6.91	182938.59	1554868.25	98	261	0%	0.01
L8_B07	25609.73	336328.95	1608441.22	164	254	1%	0.01
S2_B02	925.00	3009.29	10935.04	82	296	6%	0.00
S2_B03	1776.63	6800.98	17592.46	88	293	7%	0.00
S2_B04	1132.97	23981.59	48125.54	114	274	2%	0.00
S2_B05	2786.46	10843.44	45228.80	78	270	5%	0.00
S2_B06	5905.82	24656.34	13903.58	87	544	13%	0.00
S2_B07	7851.98	32302.77	27878.68	91	1124	12%	0.00
S2_B08	1346.73	45365.26	110942.09	91	1588	1%	0.00
S2_B11	8491.74	55456.92	293189.70	106	299	2%	0.00
S2_B12	12047.52	57790.91	291291.64	79	288	3%	0.00

Ranges given in meters.

Table 14. Parameters of fitted variogram models for Samson 2019

Variable	Nugget	Sill 1	Sill 2	Range 1	Range 2	Nugget:Sill	MSS
P	441.31	141.47	1351.68	145	564	23%	0.00
K	2189.98	1686.68	2537.14	244	466	34%	0.01
Soil pH	0.01	0.11	0.07	56	415	7%	0.01
EC_Shallow							
EC_Deep							
L7_B01	291.65	2983.20	14628.29	129	258	2%	0.00
L7_B02	0.10	10446.85	26875.80	108	260	0%	0.00
L7_B03	2274.22	29562.55	73578.34	88	326	2%	0.00
L7_B04	2324.89	75043.89	162660.72	111	552	1%	0.00
L7_B05	7251.99	117756.01	322569.05	91	299	2%	0.00
L7_B07	11379.82	116042.52	342489.67	94	311	2%	0.00
L8_B02	47.54	14135.94	13204.78	234	284	0%	0.01
L8_B03	1032.41	10137.25	26945.56	94	299	3%	0.00
L8_B04	6293.62	19458.14	80808.73	119	307	6%	0.00
L8_B05	7931.29	87477.60	100100.31	74	448	4%	0.00
L8_B06	22431.74	88732.98	277905.08	112	307	6%	0.00
L8_B07	2152.08	173869.69	193422.84	237	288	1%	0.00
S2_B02	2310.92	2958.79	8699.83	86	275	17%	0.00
S2_B03	5219.63	7706.65	14990.91	61	265	19%	0.00
S2_B04	16473.73	30263.84	16504.36	72	338	26%	0.00
S2_B05	2.87	21669.88	31216.08	50	283	0%	0.00
S2_B06	5653.08	31659.77	15831.81	60	324	11%	0.00
S2_B07	5336.09	38201.72	17242.46	57	300	9%	0.00
S2_B08	14031.83	40054.85	16661.45	79	275	20%	0.00
S2_B11	0.01	298167.83		214		0%	0.01
S2_B12	28.03	286977.21	58951.89	222	271	0%	0.01

Ranges given in meters.

FIGURES

Figure 1. Study Site Locations

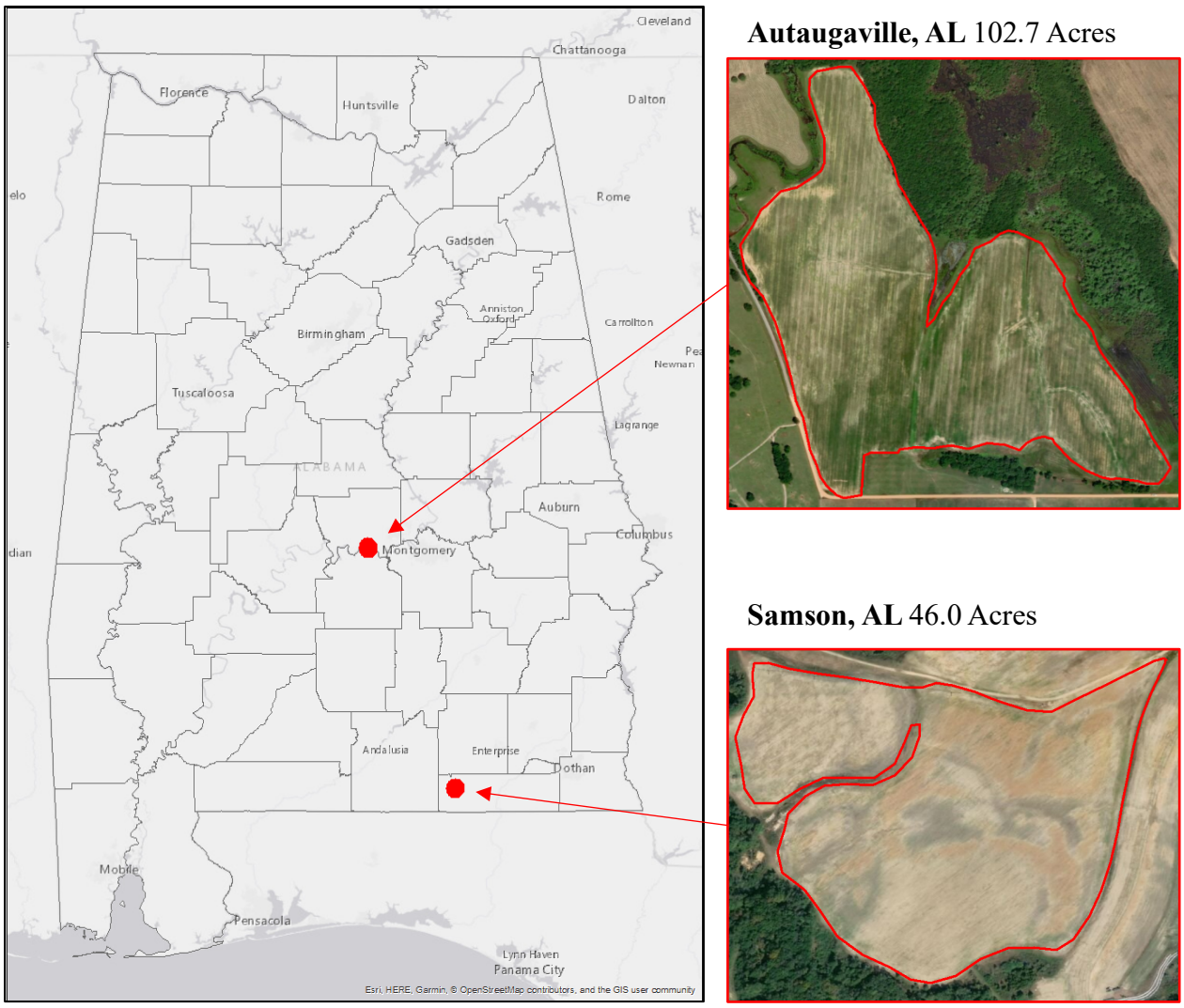


Figure 2. Study Site Soil Series (SSURGO

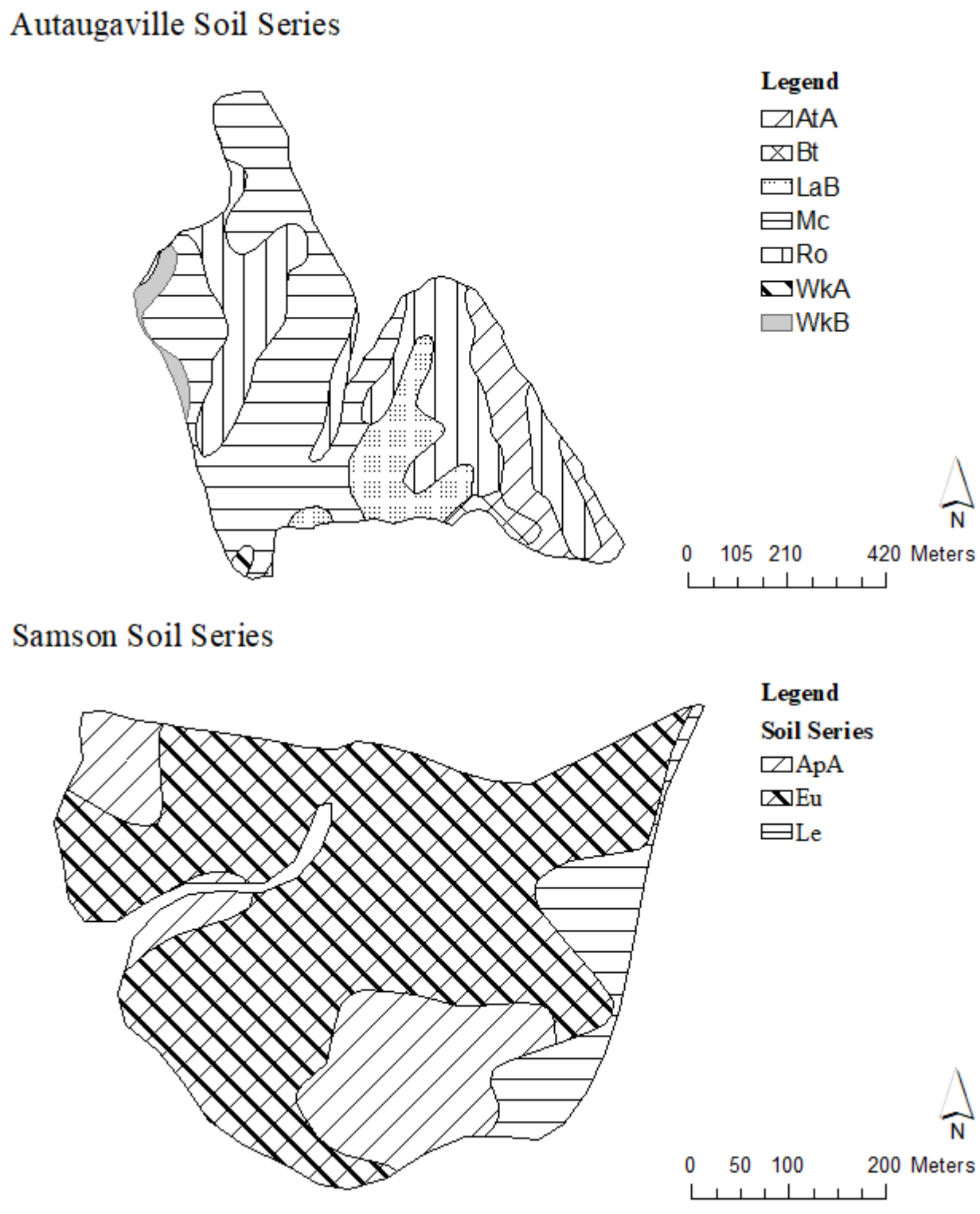


Figure 3. Soil Sampling Schemes (50 m Grid) and Interpolation Polygon (10 m Grid)

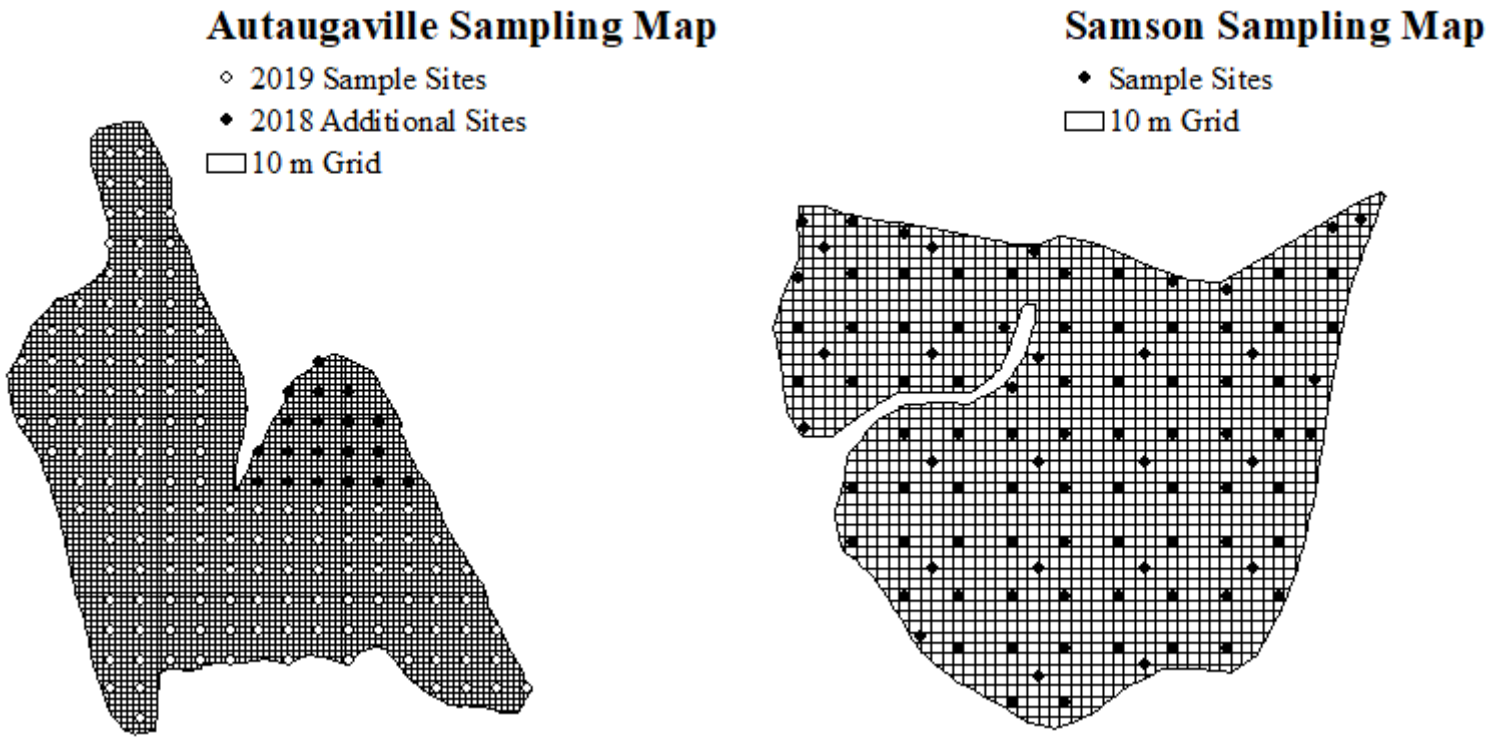


Figure 4. Shallow (1 ft) and Deep (3 ft) Soil EC (Autaugaville and Samson)

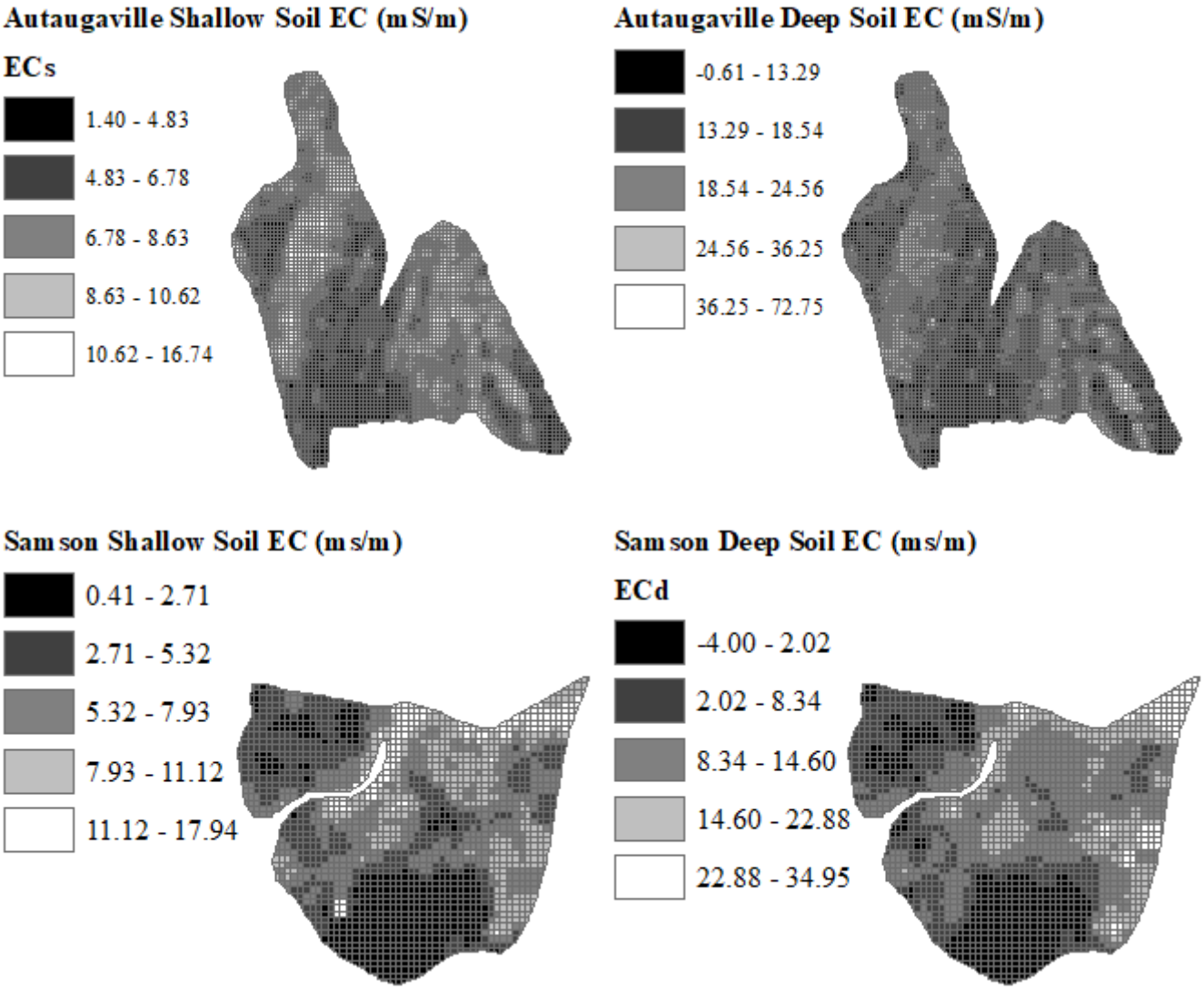


Figure 5. Autaugaville Soil Fertility Distributions

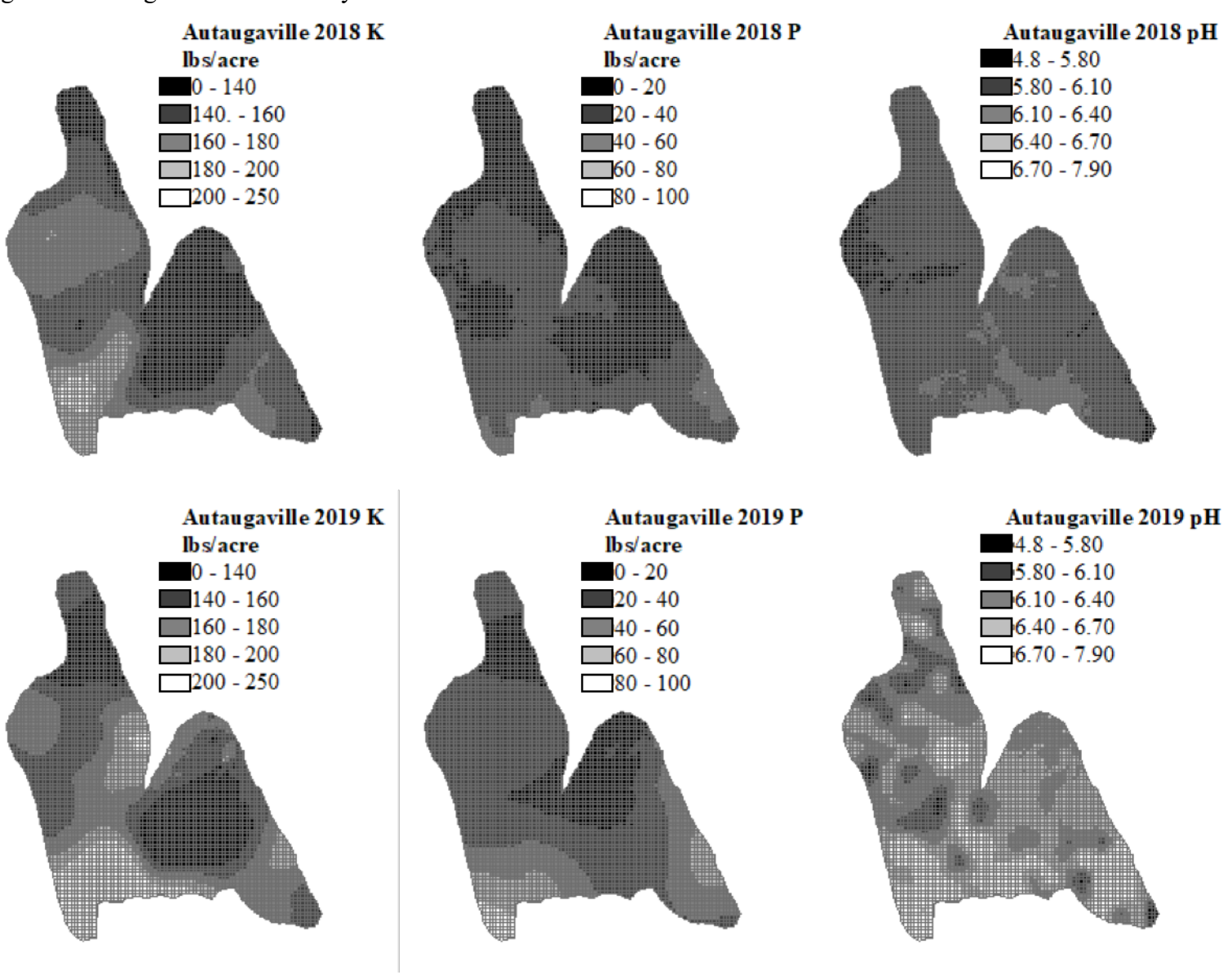


Figure 6. Samson Soil Fertility Distributions

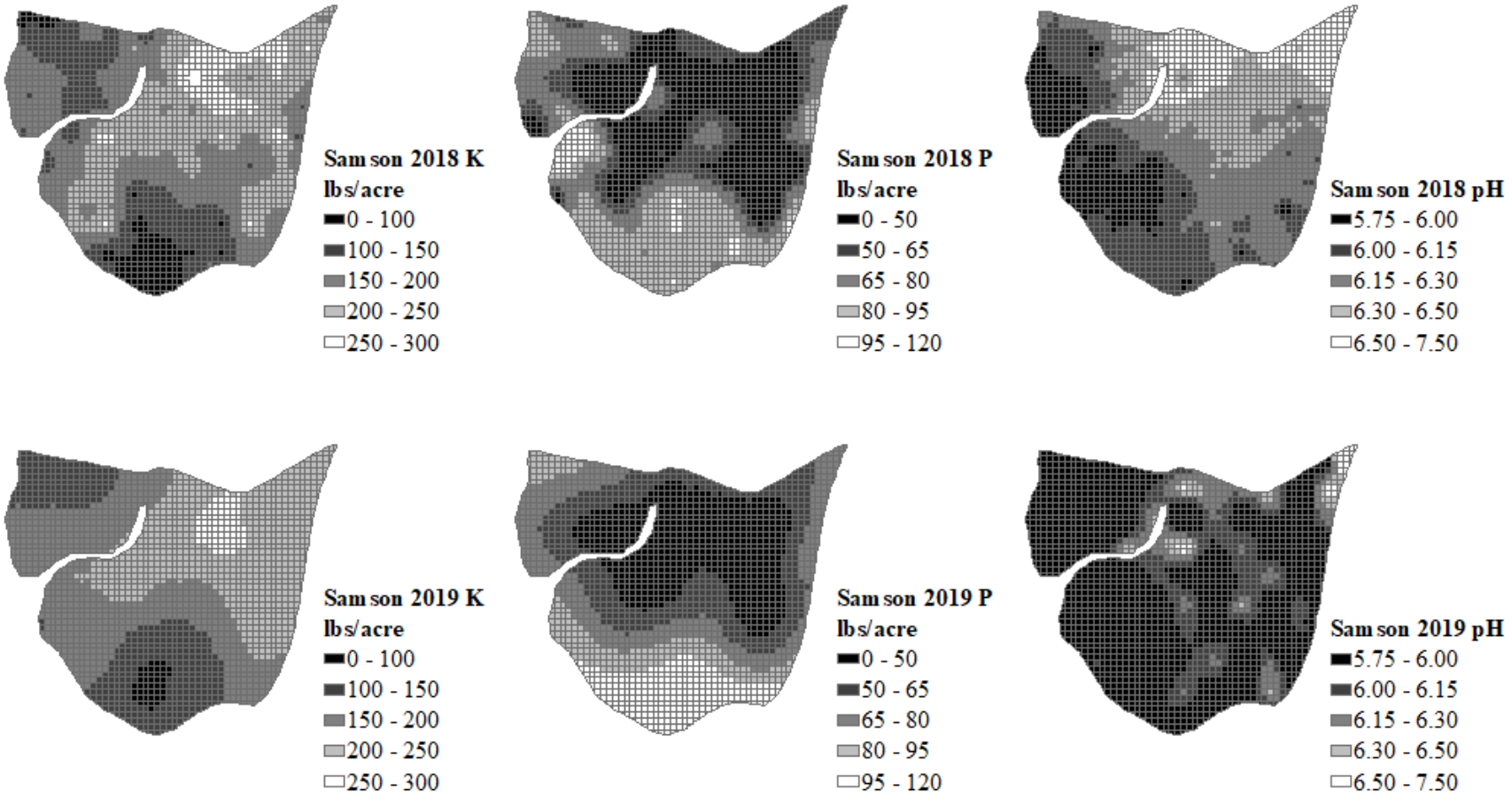
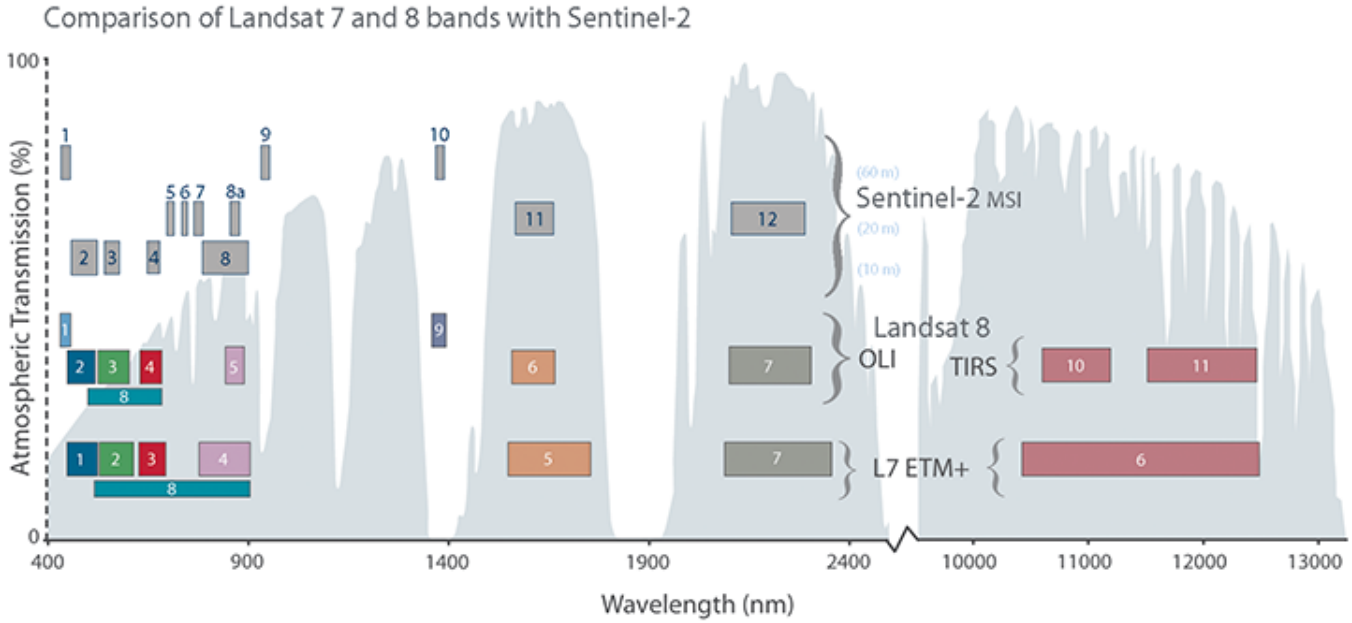


Figure 7. Remotely Sensed Data Band Relationships
Allows for the comparison of bands from the three satellites (Landsat 7, Landsat 8, and Sentinel 2) by plotting them according to their spectrum in nm.
From https://landsat.usgs.gov



LITERATURE CITED

Allington, G. 2010. Variograms - Modeling & Interpolating Spatial Dependencies. http://phylodiversity.net/azanne/csfar/images/2/28/Variograms.pdf (accessed 17 October 2019).

Ben-Dor, E., and A. Banin. 1995. Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties. Soil Science Society of America 59: 364–372. doi: 10.2136/sssaj1995.03615995005900020014x.

Bray, R.H., and L.T. Kurtz. 1945. Determination of Total Organic and Available Forms of Phosphorus in Soils. Soil Science 59(1): 39.

Ferguson, R.B., and G.W. Hergert. 2000. Soil Sampling for Precision Agriculture. University of Nebraska Extension EC 00-154: 4.

Ford, D. 2010. The Empirical Variogram. http://faculty.washington.edu/edford/Variogram.pdf (accessed 2 October 2018).

Godwin, R.J., and P.C.H. Miller. 2003. A Review of the Technologies for Mapping Within-field Variability. Biosystems Engineering 84(4): 393–407. doi: 10.1016/S1537-5110(02)00283-0.

Haynes, R.J., and R. Naidu. 1998. Influence of lime, fertilizer and manure applications on soil organic matter content and soil physical conditions: a review. Nutrient Cycling in Agroecosystems 51(2): 123–137. doi: 10.1023/A:1009738307837.

Kerry, R., and M.A. Oliver. 2008. Determining nugget:sill ratios of standardized variograms from aerial photographs to krige sparse soil data. Precision Agriculture 9(1–2): 33–56. doi: 10.1007/s11119-008-9058-0.

López-Granados, F., M. Jurado-Expósito, S. Atenciano, A. García-Ferrer, M.S. de la Orden, et al. 2002. Spatial variability of agricultural soil parameters in southern Spain. Plant and Soil 246(1): 97–105. doi: 10.1023/A:1021568415380.

Matheron, G. 1963. Principles of geostatistics. Economic Geology 58(8): 1246–1266. doi: 10.2113/gsecongeo.58.8.1246.

Mehlich, A. 1953. Determination of P, Ca, Mg, K, Na, and NH4. North Carolina Soil Test Division.

Mehlich, A. 1984. Mehlich 3 soil test extractant: A modification of Mehlich 2 extractant. Communications in Soil Science and Plant Analysis 15(12): 1409–1416. doi: 10.1080/00103628409367568.

(NRCS) National Resources Conservation Service. 1998. United States Departments of Agriculture, Official Soil Series Descriptions. Available online: https://soilseries.sc.egov.usda.gov/OSD_Docs/M/MCQUEEN.html . Accessed October 31, 2019.

(NRCS) National Resources Conservation Service. 2006. United States Departments of Agriculture, Official Soil Series Descriptions. Available online: https://soilseries.sc.egov.usda.gov/OSD_Docs/R/ROANOKE.html . Accessed October 31, 2019.

Prajapati, K. 2012. The Importance of Potassium in Plant Growth - A Review. Indian Journal of Plant Sciences 1: 177–186.

Qiao, Y., and S. Zhang. 2012. Near-Infrared Spectroscopy Technology for Soil Nutrients Detection Based on LS-SVM. In: Li, D. and Chen, Y., editors, Computer and Computing Technologies in Agriculture V. Springer Berlin Heidelberg. p. 325–335

Ryan, J., H. Ibrikci, A. Delgado, J. Torrent, R. Sommer, et al. 2012. Significance of Phosphorus for Agriculture and the Environment in the West Asia and North Africa Region. Advances in Agronomy 114: 91–153. doi: 10.1016/B978-0-12-394275-3.00004-3.

Wang, M., Q. Zheng, Q. Shen, and S. Guo. 2013. The Critical Role of Potassium in Plant Stress Response. Int J Mol Sci 14(4): 7370–7390. doi: 10.3390/ijms14047370.

CHAPTER III: UNSUPERVISED CLUSTERING ANALYSIS FOR SOIL FERTILITY MAPPING USING ANCILLARY DATA

ABSTRACT

As site-specific nutrient management has grown in popularity the need for accurate soil fertility data has increased. Unfortunately, the cost of detailed soil sampling prohibits many farmers and consultants from collecting samples at the necessary resolution. It is necessary to develop techniques using easy-to-access ancillary data to guide the creation of soil sampling schemes. Data acquired from three sources Landsat 8, Sentinel 2a, and Soil Electrical Conductivity was used to analyze the predictive power of legacy data to create sampling schemes for commonly amended soil fertility variables, phosphorus (P), potassium (K), and soil pH. First, data was divided into a training set (2018 data) and a testing set (2019 data). Principal component analysis was performed on training data before clustering by K-means was used to assign classes to individual observations. By using the component loadings as a linear combination with testing data to create testing component scores, prediction by clustering analysis is made possible. This technique resulted in an average class accuracy of 67% (K), 78% (P), and 46% (pH). Changes in accuracy are possibly caused by unfavorable field conditions, poor ancillary data quality, or a weak relationship between a soil property and ancillary data. Despite these restrictions, results indicated that classification by clustering may be used to delineate modern soil fertility distributions to guide the creation of soil sampling schemes.

INTRODUCTION

The high costs associated with intensive soil sampling often prevent farmers and consultants from collecting soil fertility samples appropriate for site-specific nutrient management.

The lack of quality data can result in suboptimal seeding rates, improper variable rate irrigation, and insufficient soil amendments for precision agriculture (Ferguson and Hergert, 2000; Godwin and Miller, 2003). It is essential that the chosen sampling scheme accurately captures the spatial variation of soil fertility variables to allow for accurate kriging and prescriptions (Kerry and Oliver, 2008; Lopez-Granados et al., 2002).

The progress of soil nutrient extraction has been well documented since the first published acid-fluoride reagent extraction of phosphorus in the 1940's (Bray and Kurtz, 1945). Since this time, the Mehlich-1 and -3 processes have been adopted as standards of nutrient extraction for their ability to predict nutrient availability across a wide range of soil types (Mehlich, 1953; Mehlich, 1984). It is only within recent decades that spectrometry technology has provided multispectral sensors capable of accurately measuring color, near-infrared (NIR), and short-wave-infrared (SWIR) reflectance. The introduction of these technologies into soil fertility management has taken many forms. The in-lab spectral recording of soil samples has become a common practice and has been shown relate with soil fertility, cation exchange capacity, and organic matter (Ben-Dor and Banin, 1995; Qiao and Zhang, 2012). These relationships can be used to guide the detection and prediction of soil properties. Unfortunately, this method is limited in application as it requires the in-field collection of soil samples and specialized laboratory equipment to record calibrated soil reflectance results. Alternatively, analysis of multispectral data collected via aircraft and satellites (remotely sensed) provides opportunities for the non-invasive large-scale analysis of soil properties.

Due to the methods by which multispectral imagery is captured, autocorrelation and multicollinearity are often present among similar bands. Principle component analysis (PCA) can be used as a dimension reduction technique to mitigate multicollinearity and identify the ancillary

variables which are contributing the most variance (Agarwal et al., 2007; Boettinger et al., 2008; Demšar et al., 2103). In addition to this, principle components, as linear combinations of ancillary data, enable data with high-dimensionality to be plotted in two-dimensional space. The projection of data then allows for unsupervised classification by iterative clustering (K-means clustering). The classification of spectral imagery has been shown to be effective in prediction of soil OM and CEC (Fox and Metla, 2005; Uno et al., 2005), soil carbon (Ladoni et al., 2010), and for the delineation of management zones (Xin-Zhong et al., 2009). This study seeks to use dimension reduction (PCA) and unsupervised classification (K-means clustering) to aid in the detection and mapping of soil fertility variables (potassium, phosphorus, and pH) and to guide the creation of soil sampling schemes for precision agriculture.

MATERIALS

Description of Study Sites:

Two study sites were used for assessment of spatial variability of soil fertility using ancillary data classification. Both sites are currently used in field-crop production with corn being the primary crop. The first site, hereby called 'Samson' is an 18.62-hectare field located at 31.115815, -86.093371, in Geneva County, Alabama near the city of Samson. The second site, hereby called 'Autaugaville' is a 41.56-hectare field located at 32.391148, -86.605381, in Autauga County, Alabama near the town of Autaugaville. Table 2 records the soil series present at each site as well as the percentage of area. The location the study sites is presented in Figure 8. The maps of soil series for both sites are presented in Figure 9.

Soil Sampling:

For both the 2018 and 2019 growing seasons, samples were collected from the Samson and Autaugaville study sites before soil amendments were applied. These samples were collected on a 50 m grid and were analyzed for Mehlcih-1 extractable P, K, and soil pH. Soil P and K were selected due to their impact on crop growth, health, and yield (Prajapati, 2012; Ryan et al., 2012; Wang et al., 2013); and in consideration that they are the only two soil macronutrients to have standardized laboratory extraction methods and recommendations. Additionally, soil pH is studied due to the large influence it has over nutrient interactions, nutrient absorption, and plant health (Haynes and Naidu, 1998).

Samples were collected by bulk sampling within a 3 m radius of each grid center, with a collection depth of six inches. One hundred fifty-six soil samples were collected in Autaugaville in 2018 and 124 samples were collected in 2019. Sample size and study area was decreased in 2019 due to flood conditions in the Northeastern portion of the study site. The full sample size for 2018 data was used for spatial interpolation as this data contributed to the spatial dependence of soil variables. The Autaugaville AOI was restricted to the area sampled in the 2019 season to provide comparable spatial results between the two years. At the Samson study site, 101 samples were collected for both 2018 and 2019. Additionally, a second grid was sampled at the Samson site, at the 100 m resolution, to increase the sample count and to ensure that changes across both short and long distances were captured. Maps of the sampling strategy at both Samson and Autaugaville are shown in Figure 10. Soil fertility distribution maps are also presented in Figures 12 (Autaugaville) and 13 (Samson).

Ancillary Data:

Remotely sensed imagery was sourced using USGS Earth Explorer. Satellite data was used from USGS' and NASA's Landsat Program (Landsat 7 and 8) as well as the ESA's Copernicus Program (Sentinel 2a). Theses satellites use multispectral sensors to record multiple bands of surface reflectance values which are corrected and calibrated by their respective space agency. Imagery was selected by two primary criteria. First, imagery captured near to the date of soil sampling was given preference. As the purpose of this study is to detect the significance of remotely sensed data on the creation of soil sampling schemes, using imagery that was taken soon before soil sampling is important as it will more accurately represent the current soil conditions that will be collected by soil sampling. Secondly, imagery was selected based on the lack of obstructions (cloud cover, vegetation, etc.) that may inhibit a clear recording of soil reflectance values.

In addition to remotely sensed data, spatial correlation results were also analyzed between soil sampling results and another commonly recorded parameters, Soil EC. Soil EC was collected at both Samson and Autaugaville in 2016 using a Veris 3100 equipped with a real-time kinematic global positioning system (RTK GPS) to record Shallow EC (approximately 1 ft), Deep EC (approximately 3 ft), and elevation values. Soil EC data is mapped for Autaugaville and Samson in Figure 11.

METHODS

Data Preprocessing:

Before analysis, all geospatial data were converted to the Universal Transverse Mercator (UTM) coordinate system in the North American Datum of 1983 (NAD83), which uses meters as the standard unit of measurement. As both sites are located in Alabama, UTM zone 16N was used.

Study sites are relatively close to the standard meridian of the projection zone; therefore, the spatial data is assumed to contain minimal scale distortion and can be accurately used for spatial analysis. Geographic transformations, tabular analysis, and spatial data creation were completed using ESRI's ArcGIS 10.4.1 for Desktop. Data clipping and organization were completed using GDAL 2.3.2 from the Open Source Geospatial Foundation.

Spatial Interpolation:

To achieve a uniform spatial distribution of both soil fertility data and ancillary data, an overlay grid was created for each study site with a polygon size of 10 m x 10 m. The 10 m grids for both Autaugaville and Samson sites can be seen in Figure 10. Values for P, K, soil pH, soil EC (shallow and deep), Landsat 8 bands 2 – 7, and Sentinel 2a bands 2 – 8, 11-12 were estimated for each grid polygon by ordinary kriging interpolation. Landsat 8 band 1 and Sentinel 2a bands 1, 9, and 10 were omitted from this study as they have been designed to capture atmospheric reflectance values rather than surface reflectance. The interpolation process effectively resamples each variable into the 10 m spatial scale, resulting in a standard sample size for all variables. The polygon count was 4,387 for Autaugaville and 2,000 for Samson.

Spatial interpolation by kriging is done by assigning weights to observed values based on their distance from the location that is to be interpolated. All interpolation processes were completed using SpaceStat 4.0.21 software published by BioMedware Inc (https://www.biomedware.com/software/spacestat/). Ordinary kriging is different from other spatial interpolation methods, such as Inverse Distance Weighted (IDW), in that it will not use observed locations beyond a threshold of spatial significance, called the range, to interpolate data at unobserved locations. The range is determined by calculating the empirical semi-variogram,

which plots the Euclidean distances of observed data pairs against the observed difference between data pairs. The empirical semi-variogram is typically sorted into bins by selecting a lag size by which data pairs are arranged. A model is then fit to best capture the spatially dependent variation of the dataset which minimize residuals and prediction error. If the spatial variability of the property under study occurs at two different spatial scales, the fit of the model can be improved by the addition of a second model, allowing for each model to better fit part of the data (short and long spatial ranges), on a nested variogram model. This fitted model is then expected to accurately estimate data at unobserved locations within the spatial scale of the data.

Principal Component Analysis:

The Principal Component Analysis (PCA) is a method of dimension reduction used to reduce the variable count of large data sets, commonly performed to allow for multidimensional cluster analysis. This is done by transforming the original variables into components, which are linear combinations of the original observations. Each component is created so that it captures as much of the variability in the original data as possible, with each new component being orthogonal to its preceding counterparts. The result is a new set of uncorrelated variables which capture decreasing amounts of variation from the original data. As principal component analysis is typically conducted using data with a mean of zero and a standard deviation of 1, ancillary data was first centered and scaled. Principal components were then created using kriged data of remotely sensed imagery and soil EC, including a total of 17 original variables.

To calculate testing component scores, testing datasets were also centered and scaled. Next, each observation was assigned a score for both PC1 and PC2 by using the sum of testing values multiplied by training component loadings. If variance is distributed similarly among variables

between both training and testing datasets, it is expected that the plotting of component scores will display a similar shape and comparable classifications.

K-means Clustering Analysis:

Unsupervised classification describes a group of algorithms designed to make inferences, and predictions, from data without referring to labelled outcomes. The final classes into which data is sorted are designed entirely by the algorithm without external guidance. These methods first became popular during the early days of machine learning and continue to provide quick and computationally easy classification results. The K-means method is a type of unsupervised classification that begins with a group of randomly scattered centroids plotted among two variables representing the X and Y axes. At both the Autaugaville and Samson study sites, their respective first two principle components were used as the X and Y axes. These components were chosen because they explain the majority of the variability present in the original data set. These randomly selected centroids serve as the beginning points for every cluster. Iterative calculations then optimize the positions of the centroids until they have been stabilized. When a centroid is stabilized, it has been plotted in a position which minimizes the sum of squares for those data observations nearest to the cluster. The resulting stabilized centroids can then be used to classify new data by determining the smallest Euclidean distance from a centroid for each observation introduced, thus predicting a classification for the new observations. This capability of K-means clustering will be used to predict the classification of soil fertility by the clustering of correlated ancillary data.

Two classes were defined for each of the soil fertility variables of interest. Soil K classification was defined as “sufficient” (≥160 lb./acre) and “insufficient” (<160 lb./acre). Soil P

classification was defined as “sufficient” (≥50 lb./acre) and “insufficient” (<50 lb./acre). Lastly, soil pH classification was defined as “high” (≥6.0) and “low” (<6.0).

It should be noted that K-means clustering begins by randomly assigning centroid locations before iterating to convergence. Because of the random component, repeated clustering results may differ slightly, even when conducted with identical source data. Despite these possible differences, it is expected that centroid convergence will always produce a similar result. For each dataset that K-means clustering was conducted upon, 1,000 repetitions of the clustering were performed, and accuracy of prediction results were averaged. K means clustering was also completed using R Statistical Computing Platform.

Training and Testing Data:

In order to assess unsupervised techniques, data is first divided into a training and testing dataset. While the training data is left unaltered, the testing dataset is given class labels and then sorted into classes which the algorithm learned from the training data. This allows the measuring of categorical accuracy on previously unseen, and unlearned, data.

In order to test the effectiveness of unsupervised classification, three repetitions were conducted to predict zones of “sufficiency” and “insufficiency”. The first two repetitions were conducted using 2018 as training data, meaning this data was used to create the first two principal components and record their loadings. Data from 2019 was then transformed, using these same loadings, and assessed for the accuracy of classification. The use of prediction components created by this process were restricted to use only at their respective study sites; meaning components created from Autaugaville 2018 training data were not tested as predictors for Samson 2019. Results from these repetitions will be referred to as “Autaugaville” and “Samson” results. The last

repetition was conducted using combined data from both Autaugaville and Samson, for both 2018 and 2019, which was randomly split into a training dataset (80%) and testing dataset (20%). This final repetition of PC classification will be referred to as the "Total Random" data.

RESULTS AND DISCUSSION

Training Data PCA:

Principal component loadings for PC1 and PC2 is given in Table 16 for the Autaugaville, Samson, and Total Random training subsets. On average, these two components explain approximately 76.2% of the variation present in ancillary dataset. Variable loadings represent the correlation between the component's linear combination and ancillary data observations. These loadings can be thought of as the variable's weight upon a linear combination forming PC1 and PC2. The larger the loading value, the greater the effect of the variable upon final component scores.

For each of the three data subsets, principal components were created using only ancillary data (Landsat 8, Sentinel 2a, and Soil EC). Soil fertility data was withheld from the creation of these principal components as its inclusion would make it impossible to test the predictive power of ancillary data. For the first principal component, both the Autaugaville data and the Total Random data assigned the largest loading values to bands from the Landsat 8 satellite, indicating that this data includes the majority of the variance present in both datasets. Contrastingly, the Samson PCA assigned the greatest loading values to the Sentinel 2a imagery. For each of the three datasets (Autaugaville, Samson, Total Random), EC shallow and EC deep were both assigned relatively weak PC loadings, which can be interpreted as a lack of correlation with remotely sensed data and therefore a lack of contribution to PC1. The wavelengths of Sentinel 2a were included in

PC2 and had most of the remaining variance from all three datasets. The domination of PC1 by Landsat 8 data and PC2 by Sentinel 2a data indicates that the use of the first two principle components will capture the majority of the variance present among all remotely sensed band with only two variables. In fact, total variance captured in Autaugaville was 40.1% for PC1 and 28.7% for PC2, a total of 68.8%. Total variance captured in Samson was 64.6% for PC1 and 15.6% for PC2, a total of 80.2%. Lastly, total variance captured in the Total Random data (Random subset from both sites during both years) was 56.6% for PC1 and 23.0% for PC2, a total of 79.6%.

Training data PCA scores, plotted by PC1 and PC2, are given by dataset. Autaugaville data is presented in Figure 14, Samson data is presented in Figure 15, and Total Random data is presented in Figure 16. Thresholds of sufficiency were set at 50 lb./acre for Phosphorus and 160 lb./acre for Potassium. The threshold for pH was set at 6.0. Visual analysis of the component score plots allows for the detection of clusters of sufficiency/insufficiency, with the exception of Autaugaville 2018 phosphorus in which nearly every sample was recorded to have sufficient values. K-means will be used to locate the centroid of these clusters which will then be used for prediction of sufficiency levels in the testing datasets.

Testing Data PCA Scores:

Testing data principal component scores created by the linear combinations created from training data are given by dataset. Autaugaville data is presented in Figure 14, Samson data is presented in Figure 15, and Total Random data is presented in Figure 16. It should be noted that in the central portion of each testing scores plot, the distribution had a similar shape to the corresponding training dataset plot. This similarity indicates that clusters from training data which correspond with this center region can be used to predict cluster properties testing datasets. A

glance at the scale of each axis will also reveal that testing distributions exhibit a dramatic increase in spread across the plane. These extreme scores are expected as testing data was not used to create component loadings. Clusters outside of the scale of training data may indicate a difference in distribution of variance or observations with relatively extreme values. It is from this difference of scale that the majority of prediction errors are expected to originate as unique clusters in this region will be ignored and will be classified by the nearest training centroid.

K-means Clustering:

While each iteration of K-means classification may give a slightly different result, due to random placement of original centroids, it is expected that every repetition will yield similar results; therefore, the results and figures given here will be averages attained over 1,000 repetitions of clustering. The number of target clusters desired is determined by observing the decrease in variance with increasing cluster count and looking for a sharp curve in the plotted variance called the 'elbow point'. This point marks the number of clusters at which variance begins to stabilize, indicating diminishing returns. The within-group sum of squares revealed a cluster count of nine optimized the reduction of variance at both Autaugaville and Samson. The elbow plots tracking variance reduction are shown in Figure 22. It was determined that nine clusters would be appropriate for classification of component scores in all three datasets. Figures 17 through 19 display the principal component scores for Autaugaville 2018, Samson 2018, and Total Random Training data sets. The nine clusters, determined by K-Means, is plotted at the top of each figure, followed by the classification of each dataset. The classifications for each cluster were determined by the class majority within the cluster.

It should be noted that for Autaugaville potassium and phosphorus, and Samson pH, each cluster received the same classification, although observations within each cluster were not homogenous. This is a result of thorough "mixing" of soil fertility observations among PC1 and PC2. During prediction, these datasets will be represented as a uniform surface with only the majority classification being assigned. In these instances, K-means classification of the first two principal components is not a suitable methodology, as it yields predictions that are only as accurate as an assumption of within-field uniformity.

Fortunately for Autaugaville pH, Samson potassium and phosphorus, and every Total Random dataset variable, the classification assignments are visually similar to the observed values from training datasets (as can be seen through comparison distribution shapes in Figures 14, 15, and 16). In addition to the similar "shape" of training and testing component scores, this visual similarity further indicates that principal component clustering may be used to predict sufficient/insufficient areas of soil fertility variables.

Clustering Accuracy Results:

Results for using K-means classification as a means of categorical prediction in soil fertility variables is given in Table 17. Maps of predicted observation classes are presented in Figure 20 (Autaugaville) and 21 (Samson). These results were attained by first calculating that majority class of each centroid fitted on training data, meaning that if the majority of the observation associated with a centroid were classified as "sufficient", then the centroid itself was classified as "sufficient". These centroids, which originated from the principal components of training data, were then plotted upon the PC1 and PC2 from the testing data. A Euclidean distance from each centroid to every observation was calculated, and observations were classified as sufficient/insufficient

according to the class of their nearest centroid. Accuracy results are given as a percentage of observations in the test datasets which were correctly classified as sufficient or insufficient.

On average, soil potassium (67% accuracy) and soil phosphorus (78% accuracy) outperformed the predication of soil pH (46% accuracy). This is also unsurprising, considering the relatively weak correlation coefficients observed between soil pH and remotely sensed ancillary data in Chapter II. As a whole, unsupervised classification of aerial imagery is best fitted for use in predicting zones of sufficiency in soil phosphorus, with limited applications in soil potassium. It should be noted that the Total Random dataset has the most consistent accuracy which is to be expected as is it only $1/4^{th}$ the size of its corresponding training dataset, compared to data from Autaugaville and Samson in which training and testing sets were of equal size.

CONCLUSION

By using K-means unsupervised classification it is possible to predict, with approximately 84% accuracy, zones of sufficient and insufficient soil phosphorus with nothing more than last year's soil sampling results and modern satellite imagery. Similarly, potassium shows great potential with a possible 82% accuracy in prediction of sufficiency. Yet this same method may result in relatively poor predictors, with accuracy ranging from 52% - 65%, barely better than randomly selected guesses. What is it that contributes to the variation in accuracy results?

One factor worth considering is that of image collection, namely capture date and lack of obstructions. While careful care was taken in this study to select imagery which provided the clearest view of bare soil close to the time of spring soil sampling, it is impossible to ensure that high-quality and unobstructed imagery will be available near the desired dates. It has been noted that the Samson study site employed a regular practice of tilling and spraying winter cover crops,

providing an opportune time for image capture before spring planting. On the other hand, the Autaugaville study site did not employ these same practices and included crop residue and vegetative cover all winter, through the spring planting date. This clearer and more accurate imagery likely contributed to the stronger accuracy results achieved at the Samson study site, as erroneous readings from obstructions will alter the principal component's ability to capture variance and detect patterns in soil reflectance. This is likely to limit the usefulness of remote sensing and classification techniques in modern cropping systems which encourage the use of cover crops and no-till methods.

Additionally, classification by clustering was affected largely by the distribution of the training data. When the soil fertility was relatively uniform, such as Autaugaville 2018 P and K and Samson 2018 pH, it was predicted that testing data clusters also shared this uniformity. This, in effect, masked all variation present in testing predictions limiting the use of this methodology in fields that experience extreme temporal variation of soil fertility variables. On the other hand, fields that maintain a similar soil distribution over time displayed the potential for prediction by unsupervised clustering. This prediction has the potential to delineate zones of sufficiency or insufficiency. These can be used to guide the creation of soil sampling schemes by reducing the number of samples taken in "sufficient" zones which will require fewer soil amendments.

Remote sensing classification of soil fertility variables remains an area of interest for future research. The methods used here, in particular PCA, are well suited for wide datasets. The use of hyperspectral imagery (anywhere from 100-2000 distinct bands) with this methodology could make greater use of data reduction to enable classification with only the most correlated of reflectance wavelengths. The introduction of similar training data is also likely to increase the efficiency and accuracy of this procedure. The use of multiple years of historical data have

potential to drastically increase the prediction power for an individual study site. Lastly, the collection of many datasets from one geographic region with similar soil types is also likely to enable the creation of more detailed models with greater region-specific accuracy.

TABLES

Table 15. Study Site Soil Series

Samson Soils

Soil Series	% of AOI	Taxonomic Class
Alpin Sand	20.6	Thermic, coated Lamellic Quartzipsamments
Eunola Sandy Loam	69.7	Fine-loamy, siliceous, semiactive, thermic Aquic Hapludults
Leaf-Lenoir Complex	9.7	Fine, mixed, semiactive, theric Aeric Paleaquults

Autaugaville Soils

Soil Series	% of AOI	Taxonomic Class
Altavista Loam	9.9	Fine-loamy, mixed, semiactive, thermic Aquic Hapludults
Blanton Loamy Sand	1.7	Loamy, siliceous, semiactive, thermic Grossarenic Paleudults
Lakeland Loamy Sand	10.3	Thermic, coated Typic Quartzipsamments
McQueen Silt Loam	43.7	Fine, mixed, semiactive, thermic Typic Hapludults
Roanoke Complex	32.0	Fine, mixed, semiactive, thermic Typic Endoaquults
Wickham Fine Sandy Loam	2.3	Fine-loamy, mixed, semiactive, thermic Typic Hapludults

Table 16. Principal Component Loadings

Each loading represents the weight of the individual variable upon the total score for each observation. The linear combination of these loadings and the original observations results in the Principal Component Score.

Autaugaville 2018			**Samson 2018**			**Total Random**		
	PC1	*PC2*		*PC1*	*PC2*		*PC1*	*PC2*
L8B02	-0.355	-0.113	L8B02	0.248	0.247	L8B02	0.301	0.116
L8B03	-0.358	-0.113	L8B03	0.246	0.216	L8B03	0.300	0.124
L8B04	-0.357	-0.114	L8B04	0.251	0.261	L8B04	0.298	0.129
L8B05	-0.355	-0.122	L8B05	0.055	-0.358	L8B05	0.294	0.143
L8B06	-0.351	-0.116	L8B06	0.276	0.192	L8B06	0.291	0.145
L8B07	-0.353	-0.106	L8B07	0.273	0.229	L8B07	0.291	0.148
SB02	0.229	-0.211	SB02	0.285	0.061	SB02	-0.261	0.108
SB03	0.187	-0.334	SB03	0.287	0.022	SB03	-0.277	0.183
SB04	0.221	-0.260	SB04	0.276	0.070	SB04	-0.178	0.366
SB05	0.133	-0.401	SB05	0.283	-0.064	SB05	-0.247	0.284
SB06	-0.045	-0.351	SB06	0.211	-0.387	SB06	-0.270	-0.028
SB07	-0.062	-0.329	SB07	0.196	-0.421	SB07	-0.267	-0.053
SB08	-0.025	-0.357	SB08	0.187	-0.398	SB08	-0.271	-0.003
SB11	0.130	-0.314	SB11	0.278	0.026	SB11	-0.107	0.441
SB12	0.145	-0.272	SB12	0.276	0.051	SB12	-0.060	0.473
ECs	0.151	0.074	ECs	-0.179	0.260	ECs	-0.028	-0.301
ECd	0.153	0.019	ECd	-0.194	0.193	ECd	-0.062	-0.346

Table 17. Classification Prediction Accuracy – Proportion of Cells classified correctly

Autaugaville

Potassium	*Phosphorus*	*pH*
52.06%	84.70%	33.54%

Samson

Potassium	*Phosphorus*	*pH*
82.41%	65.89%	28.28%

Total Random

Potassium	*Phosphorus*	*pH*
66.46%	83.19%	77.05%

FIGURES

Figure 8. Study Site Locations

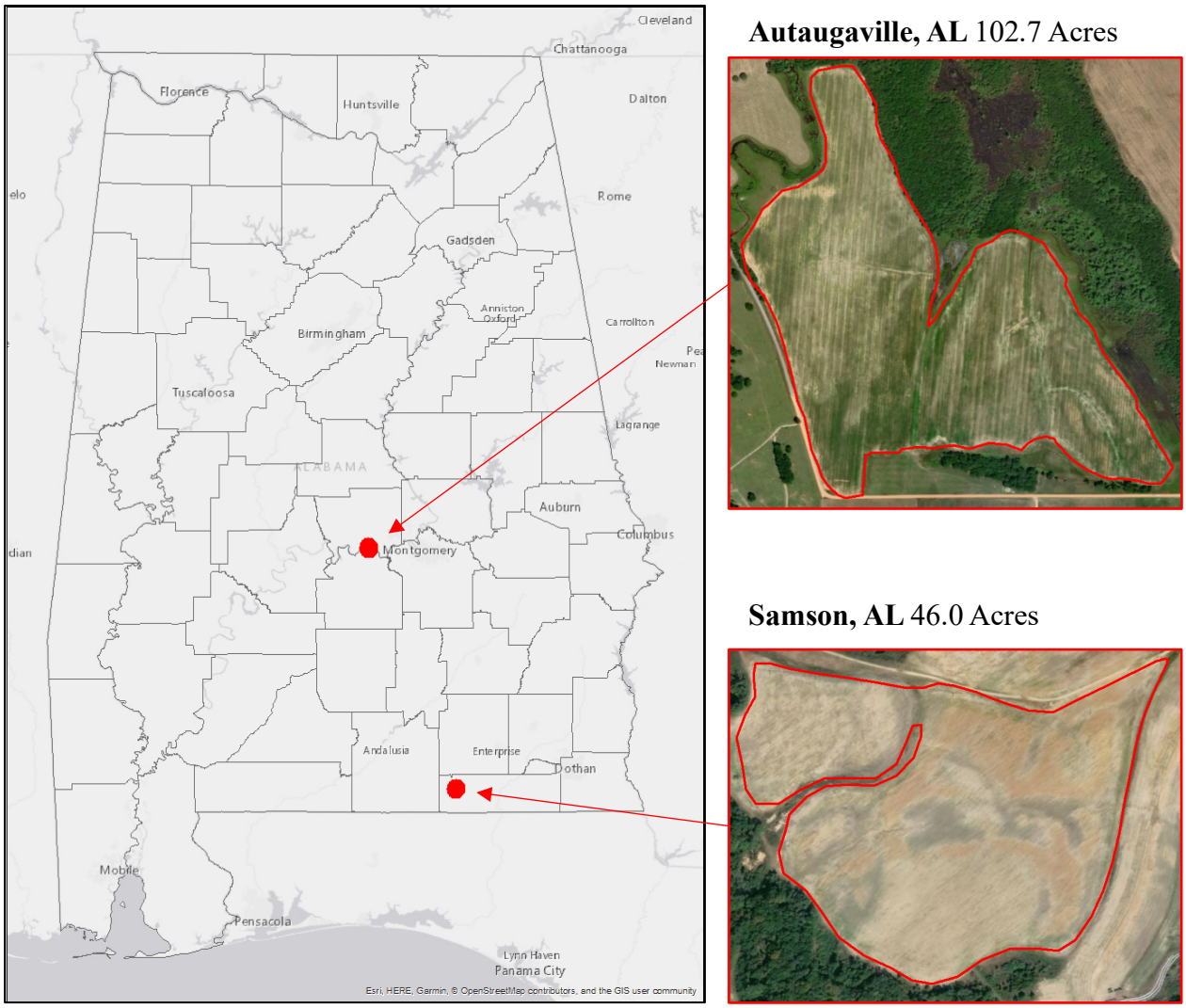


Figure 9. Study Site Soil Series (SSURGO)

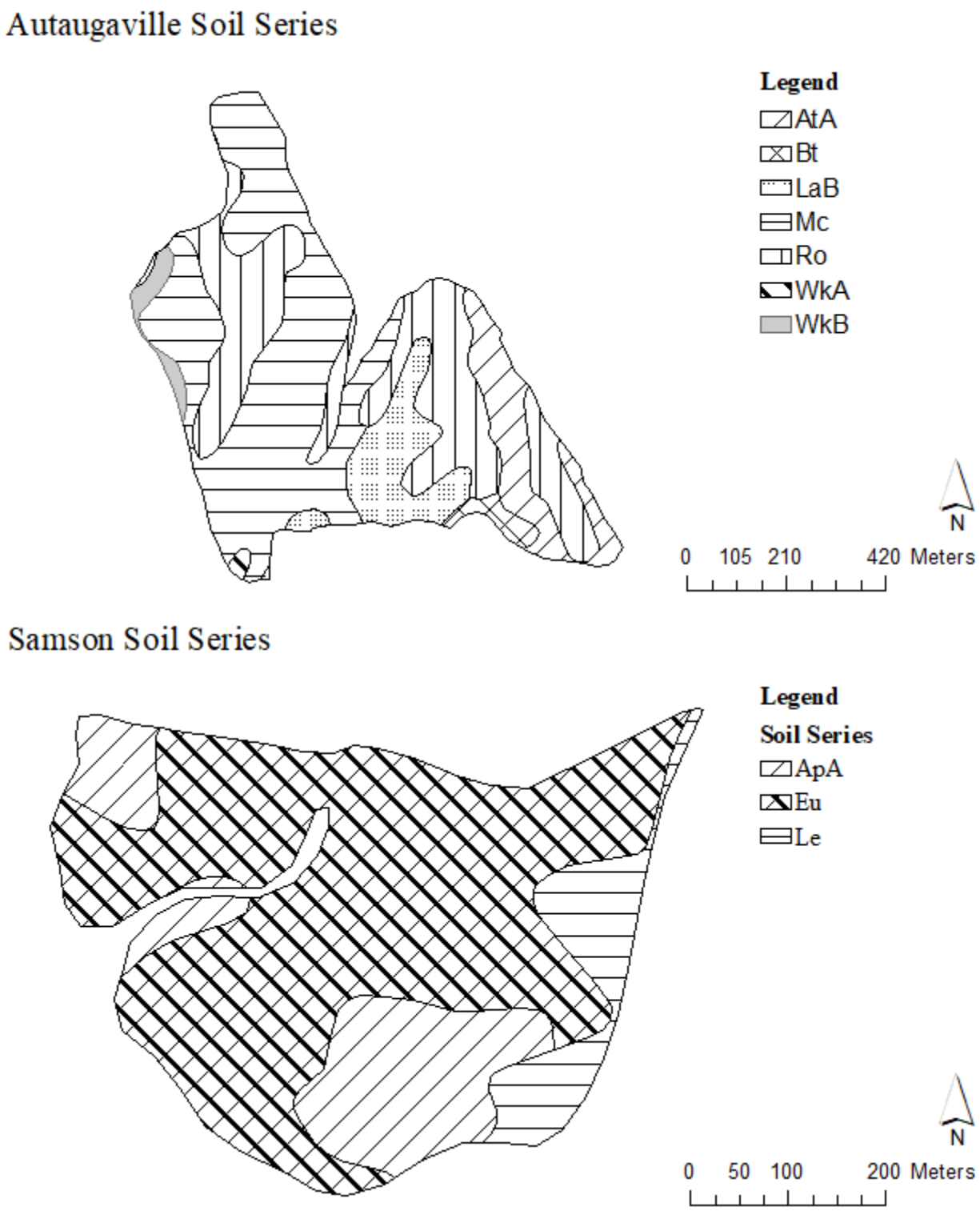


Figure 10. Soil Sampling Schemes

Both sites were sampled at a 50 m resolution. Samson study site was also sampled with an additional 100 m grid to increase sample size and better capture spatial variability of soil characteristics.

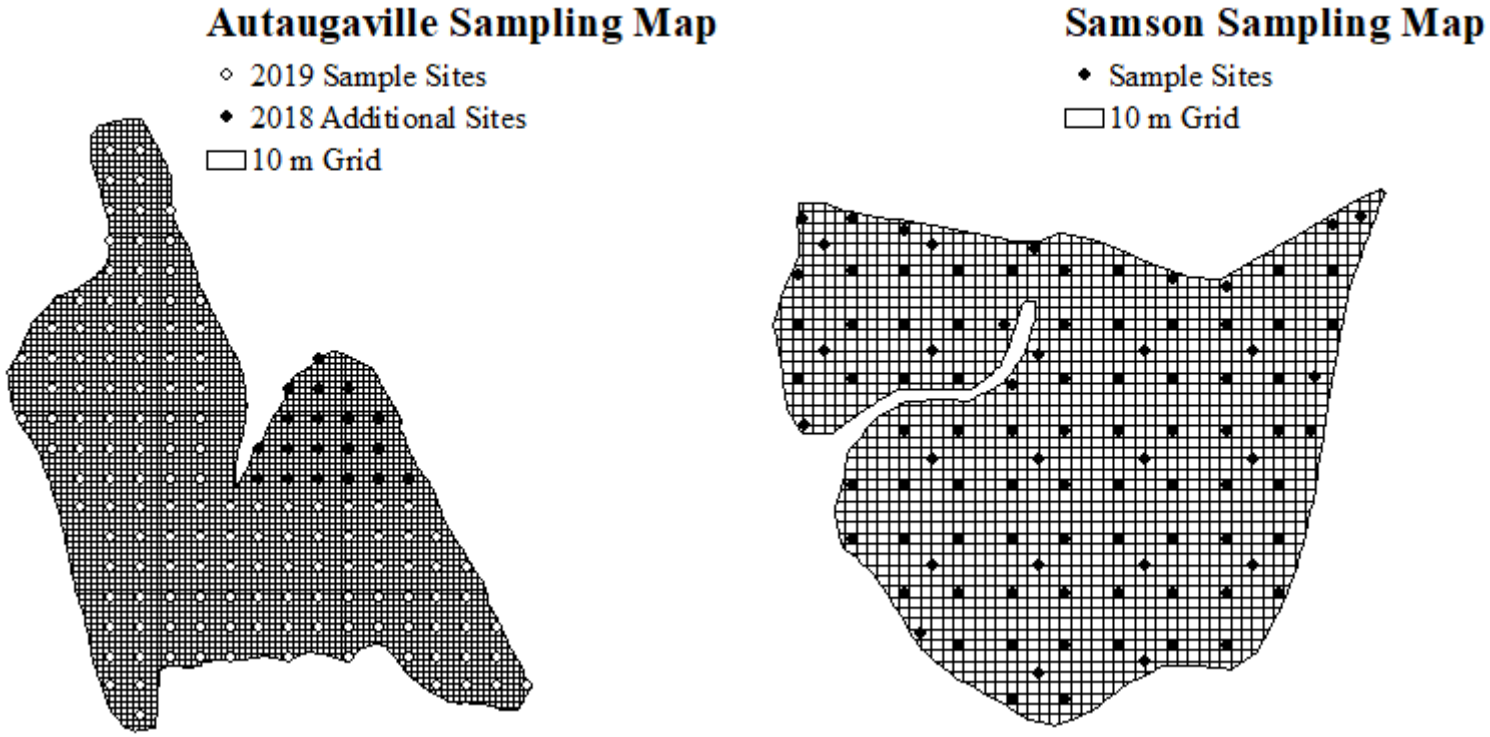


Figure 11. Shallow (1 ft) and Deep (3 ft) Soil EC (Autaugaville and Samson)

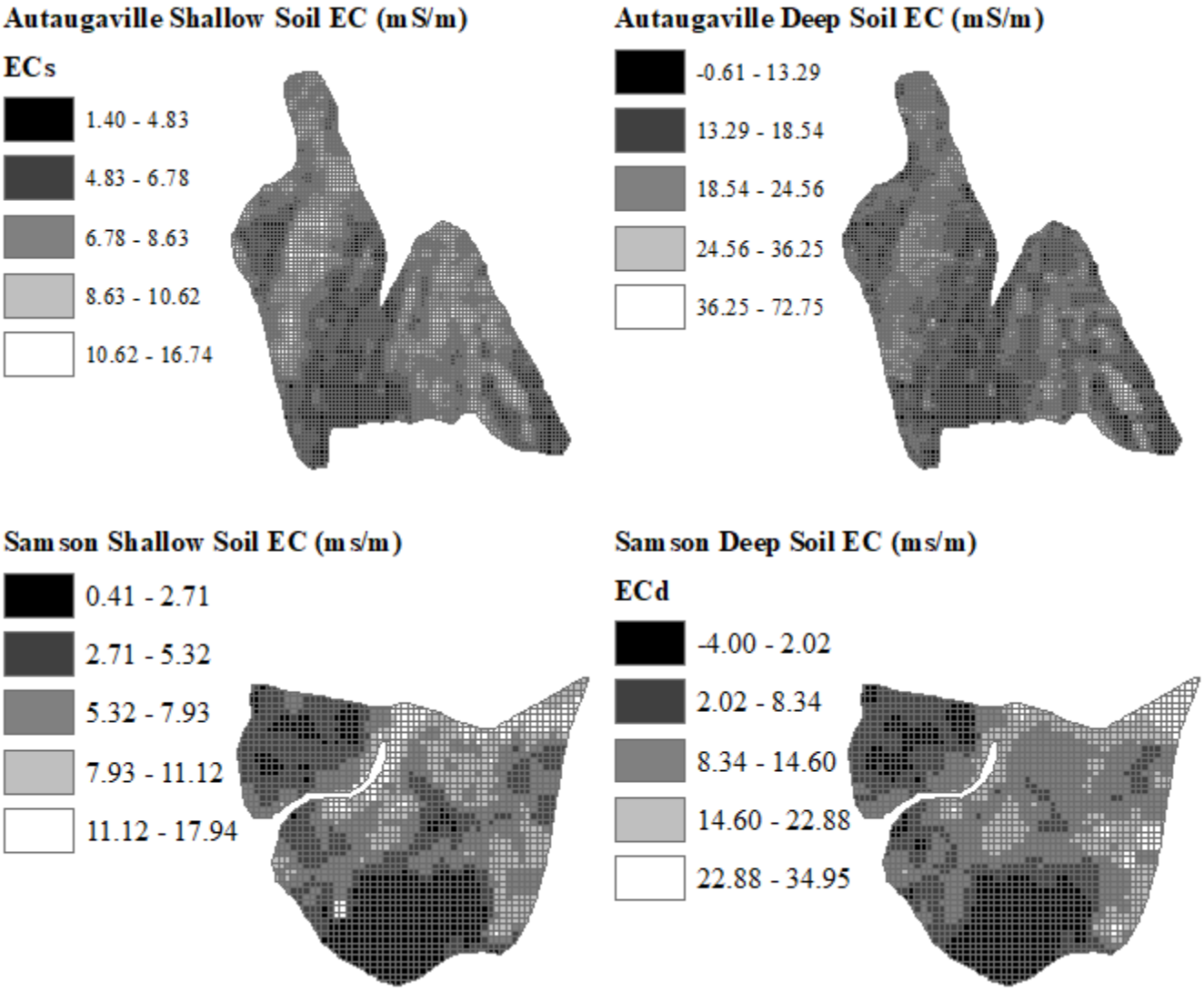


Figure 12. Autaugaville Soil Fertility Distributions

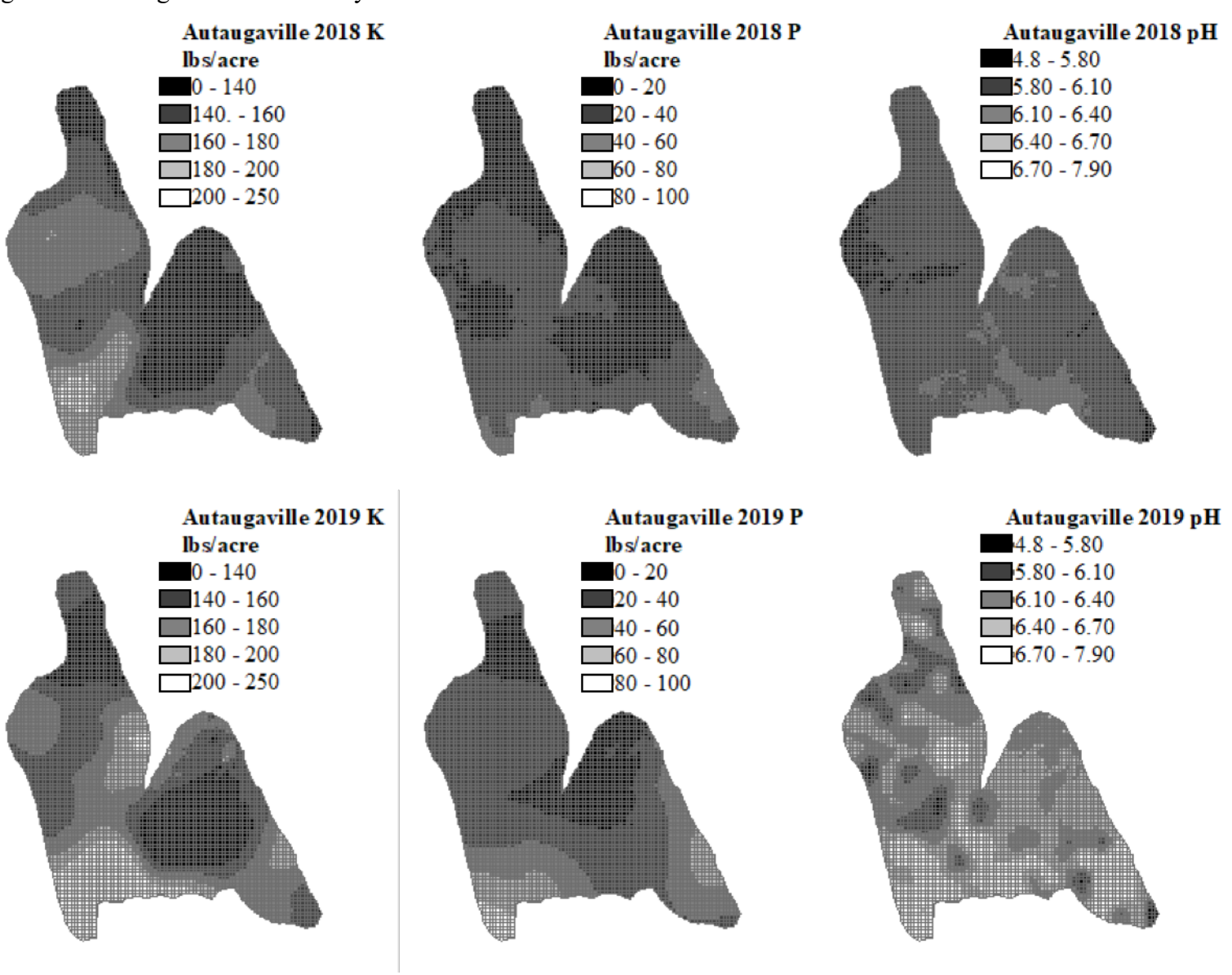


Figure 13. Samson Soil Fertility Distributions

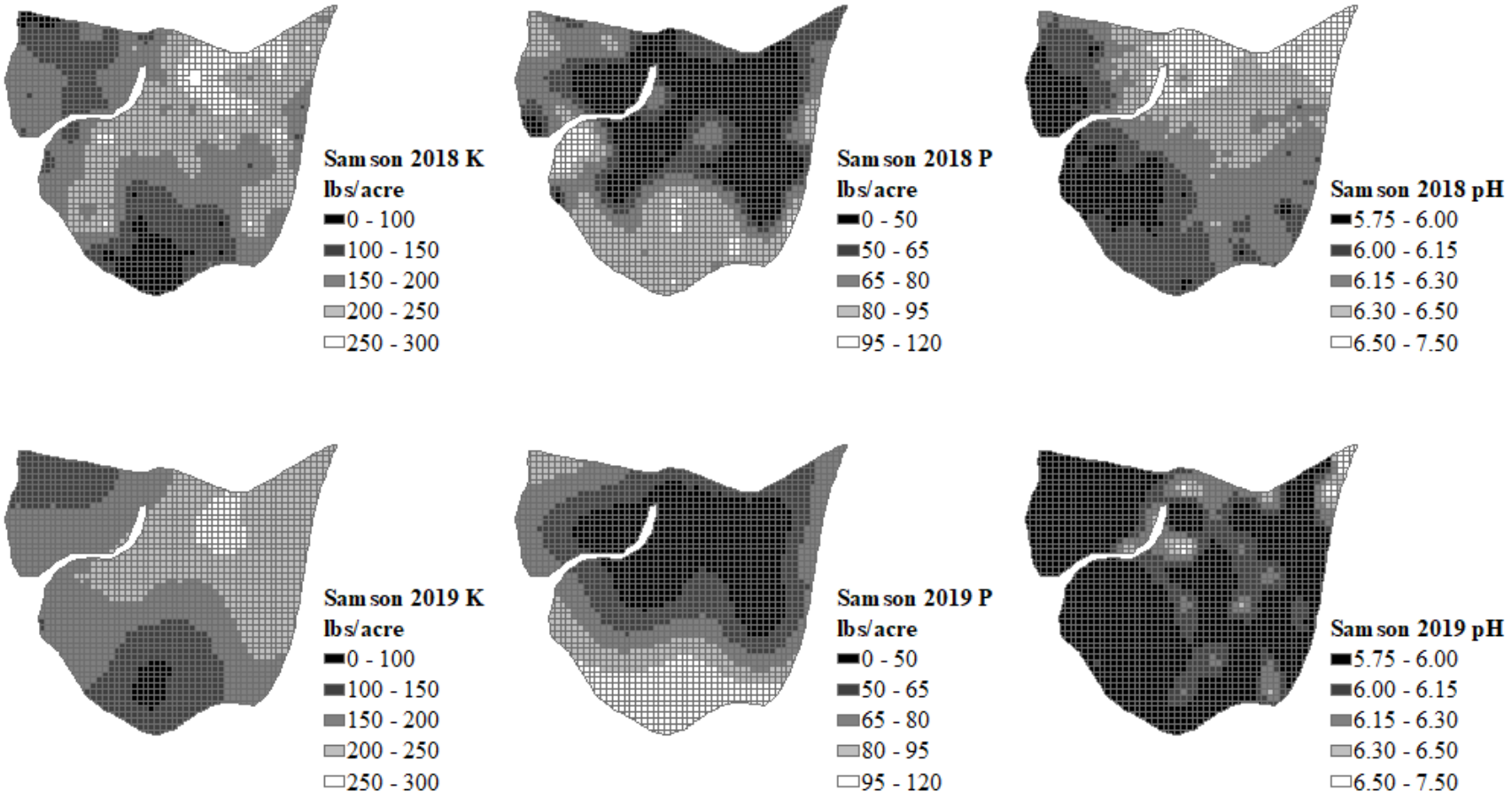


Figure 14. Autaugaville Principal Components Plots
Row 1 displays the 2018 observations plotted by PC1 and PC2. Row 2 displays the 2019 observations plotted by PC1 and PC2.

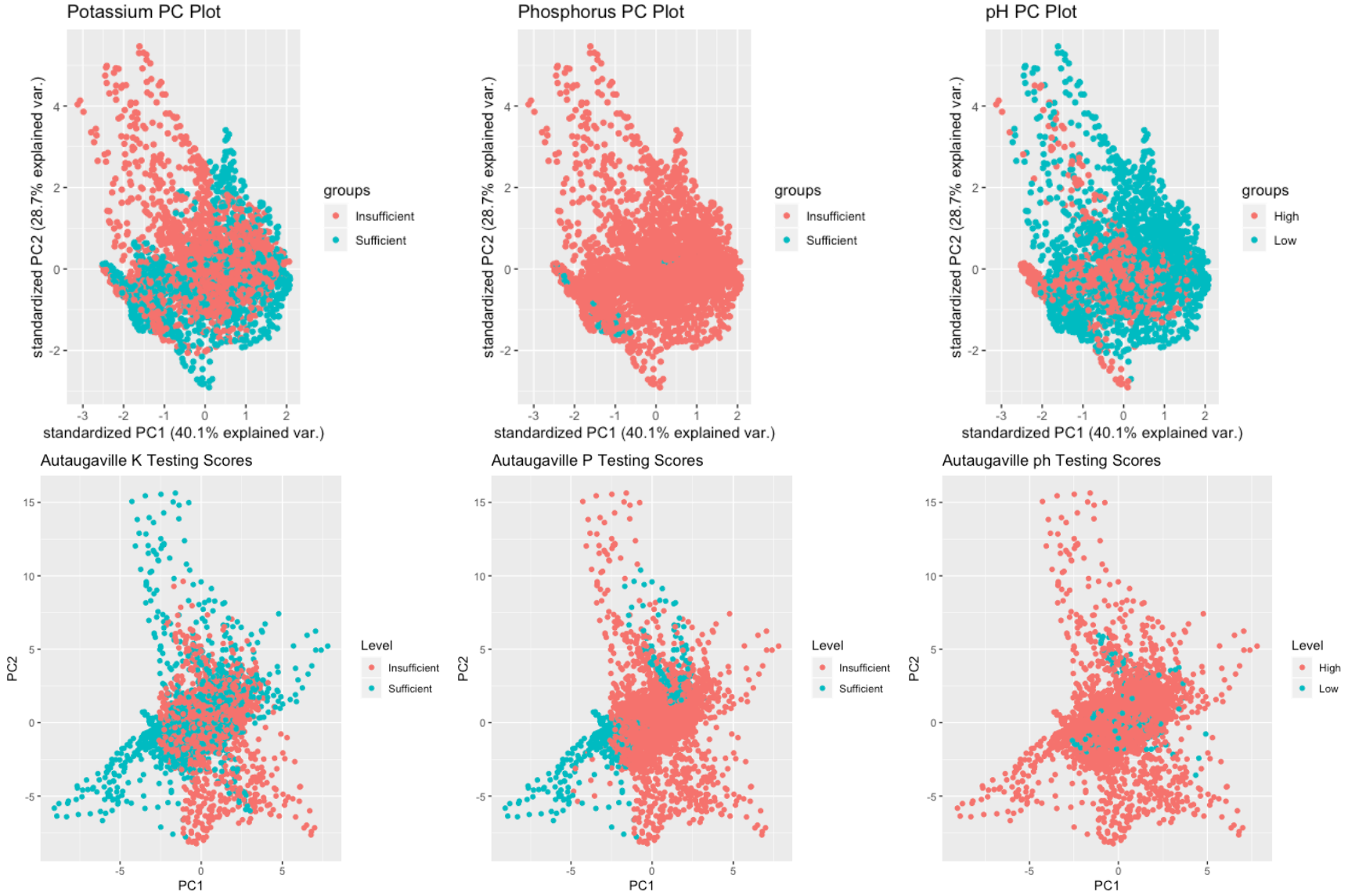


Figure 15. Samson Principal Component Plots
Row 1 displays the 2018 observations plotted by PC1 and PC2. Row 2 displays the 2019 observations plotted by PC1 and PC2.

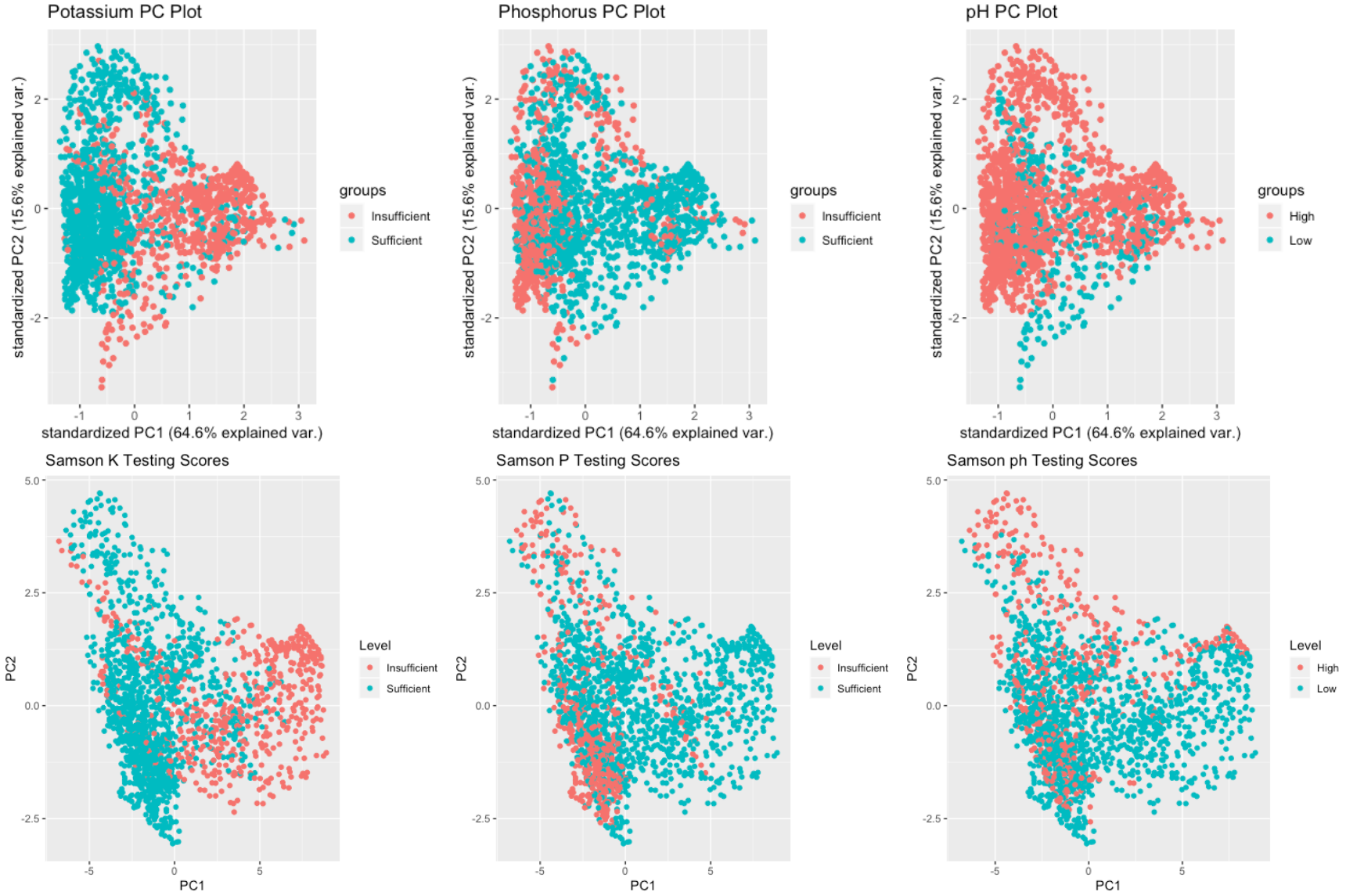


Figure 16. Total Random Data Principal Component Plots
Row 1 displays the 2018 observations plotted by PC1 and PC2. Row 2 displays the 2019 observations plotted by PC1 and PC2.

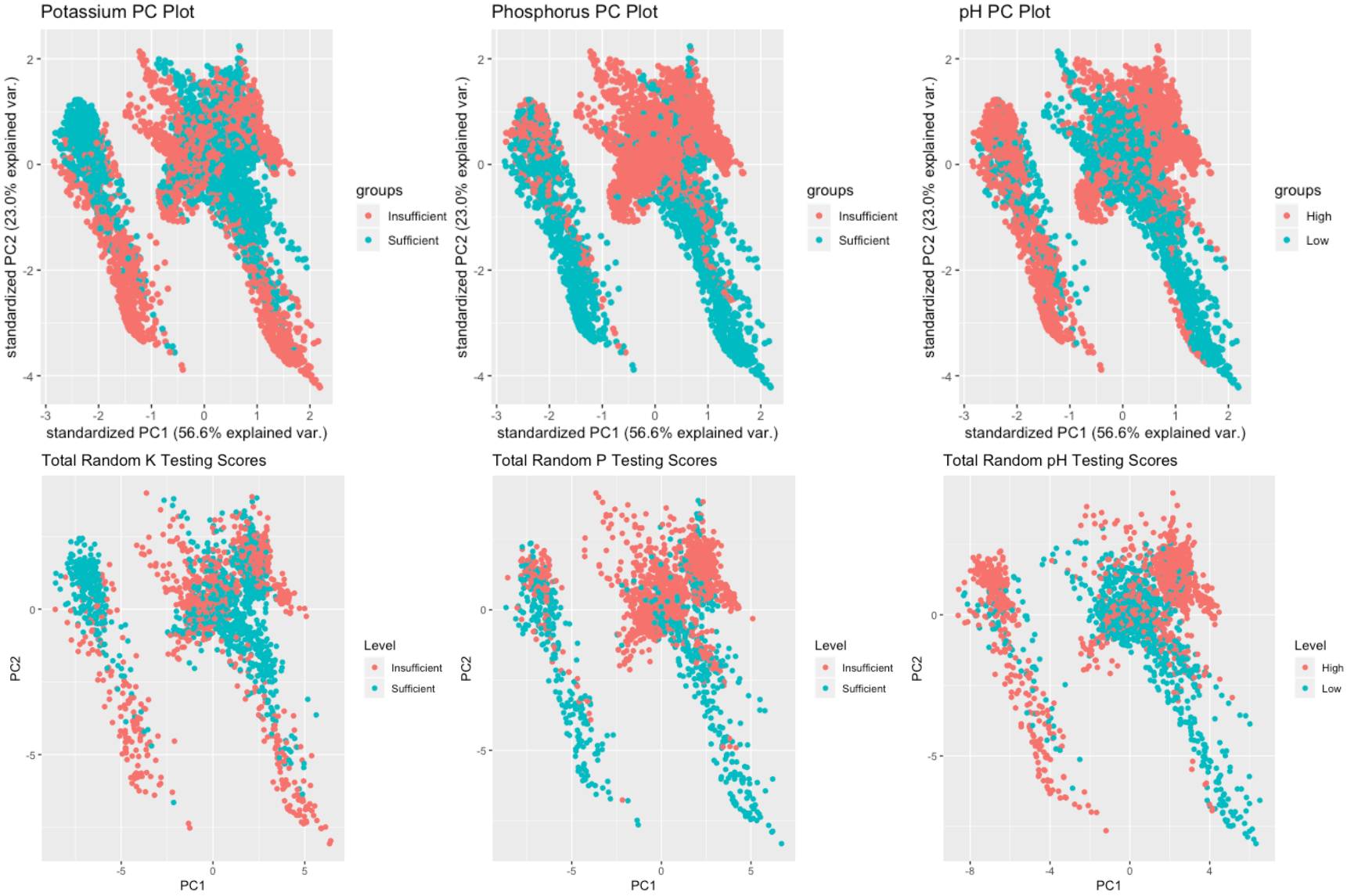
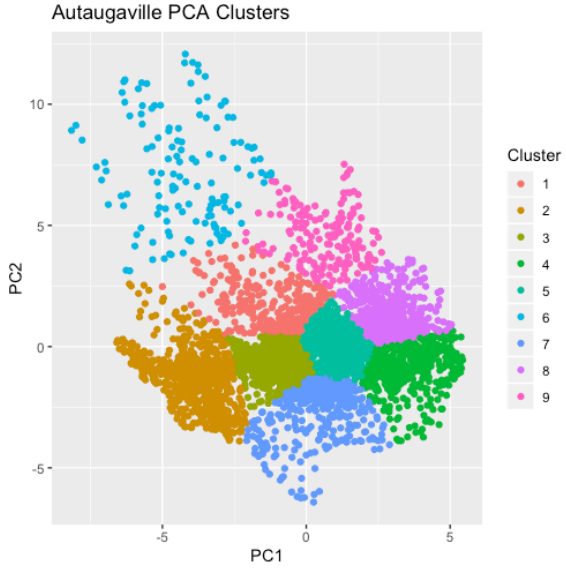
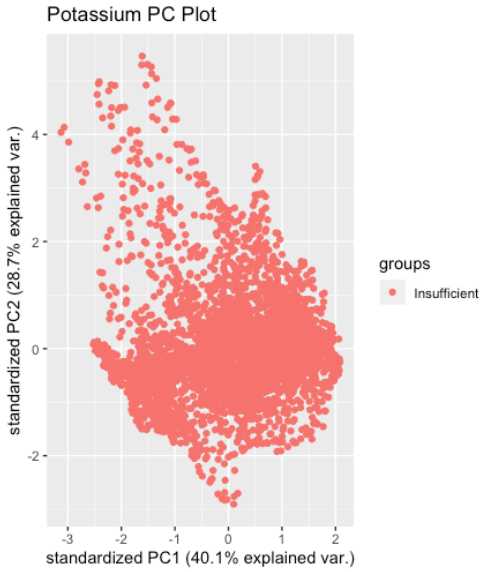
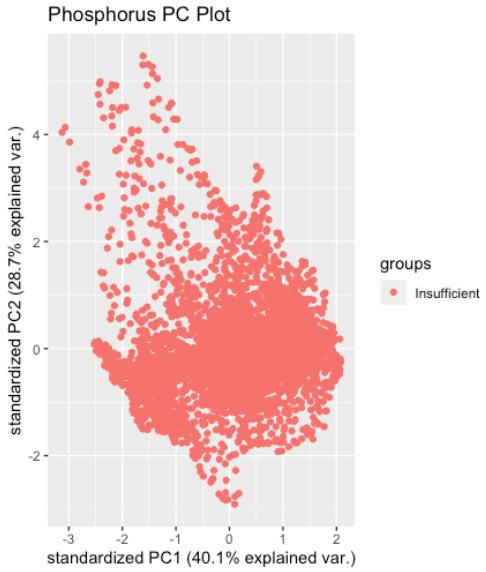


Figure 17. Autaugaville Training data (2018) clusters, and classes as assigned by class majority within cluster.







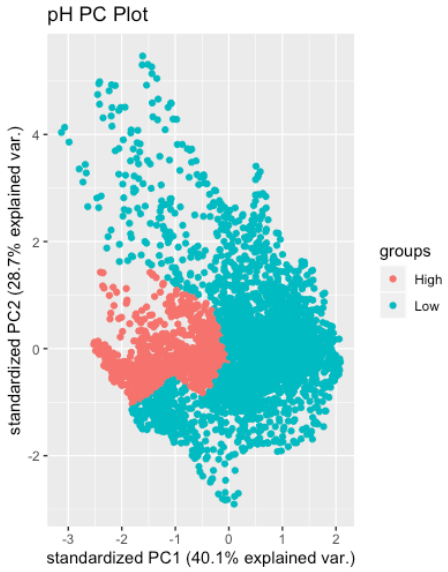


Figure 18. Samson Training data (2018) clusters, and classes as assigned by class majority within cluster.

Figure 19. Total Random training data clusters, and classes as assigned by class majority within cluster.

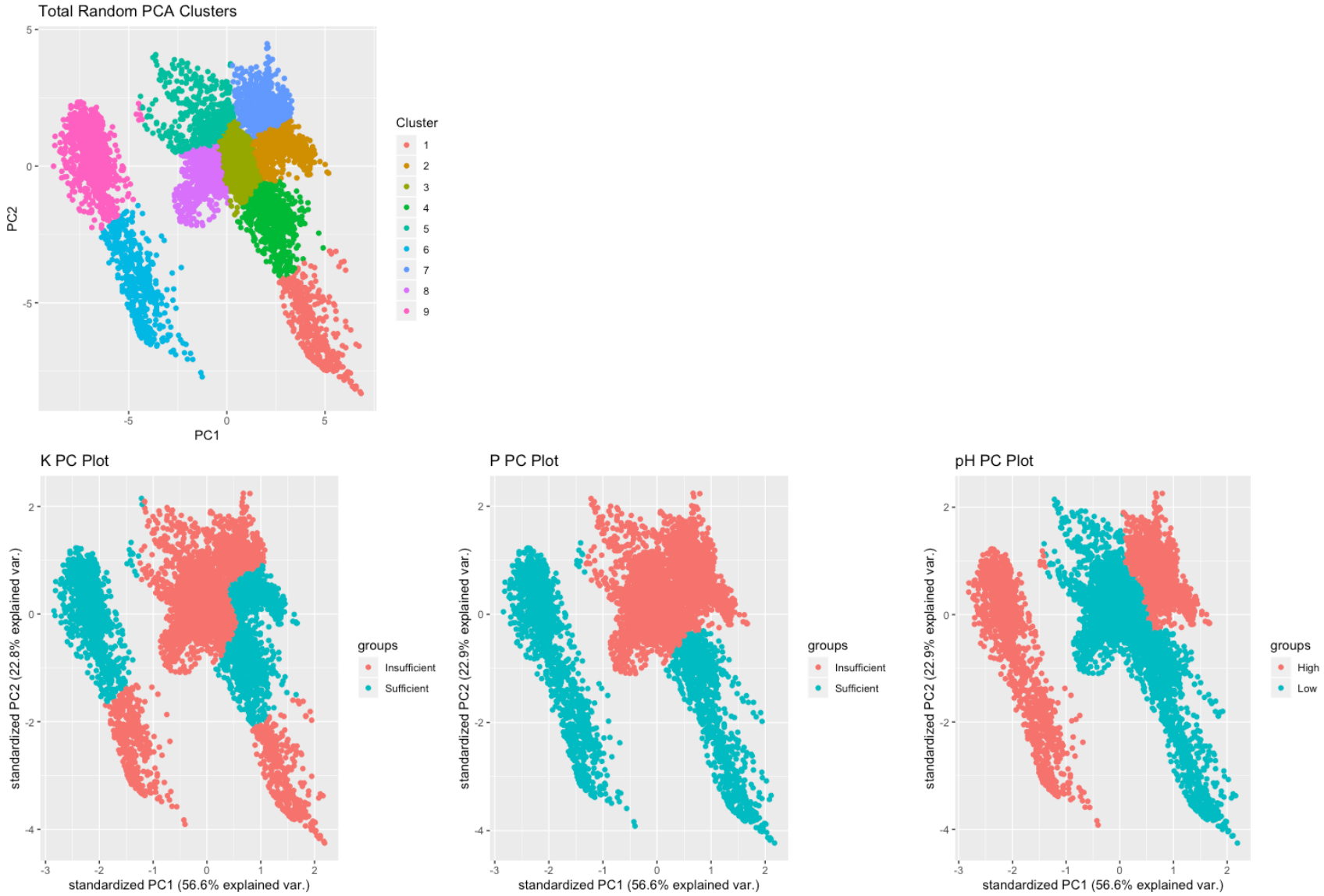


Figure 20. Autaugaville 2019 Soil Fertility Ground Truths and Unsupervised Predictions Assuming K Sufficiency at 160 lb/acre, P sufficiency at 50 lb/acre, and pH threshold at 6.0.

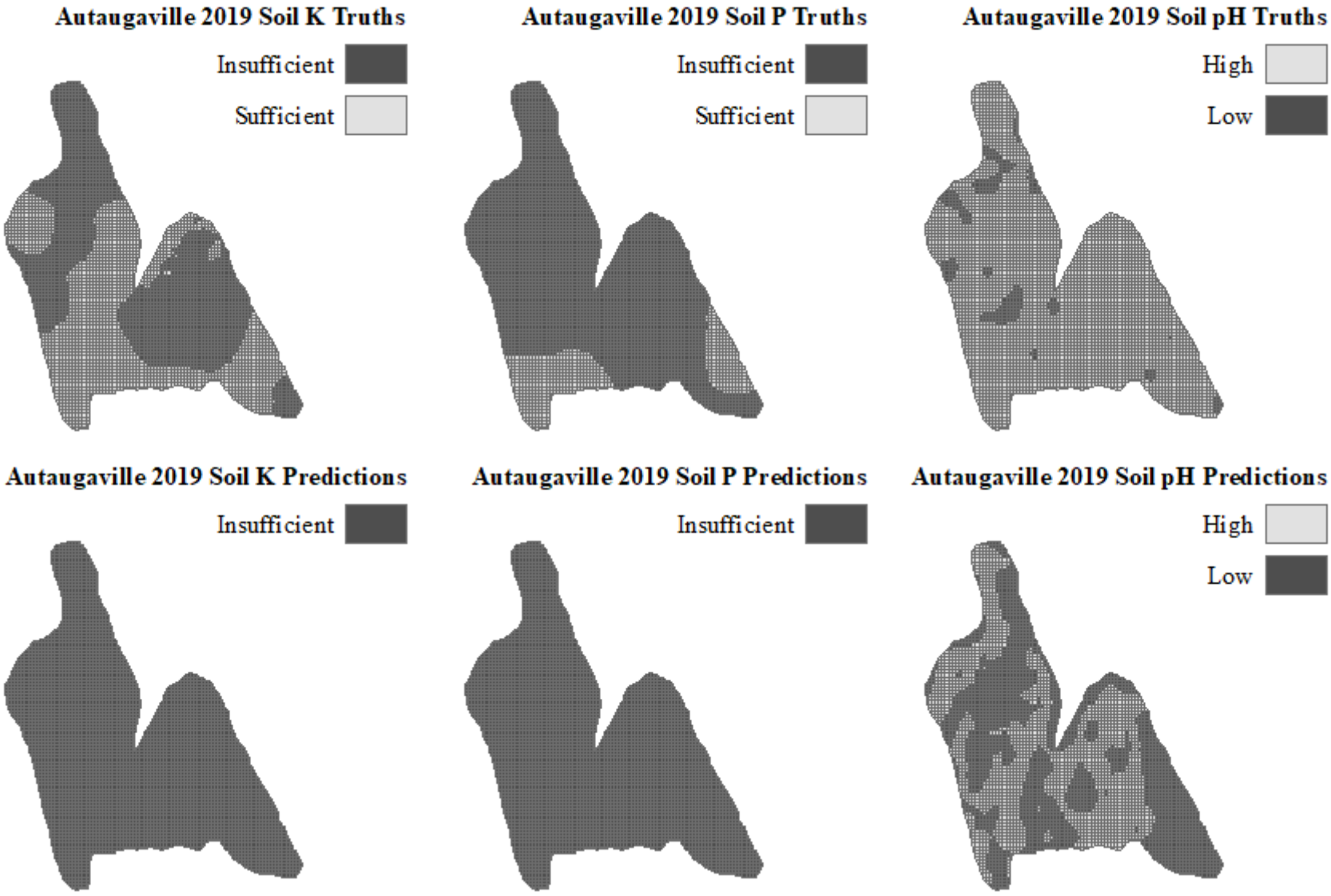


Figure 21. Samson 2019 Soil Fertility ground Truths and Unsupervised Predictions
Assuming K Sufficiency at 160 lb/acre, P sufficiency at 50 lb/acre, and pH threshold at 6.0.

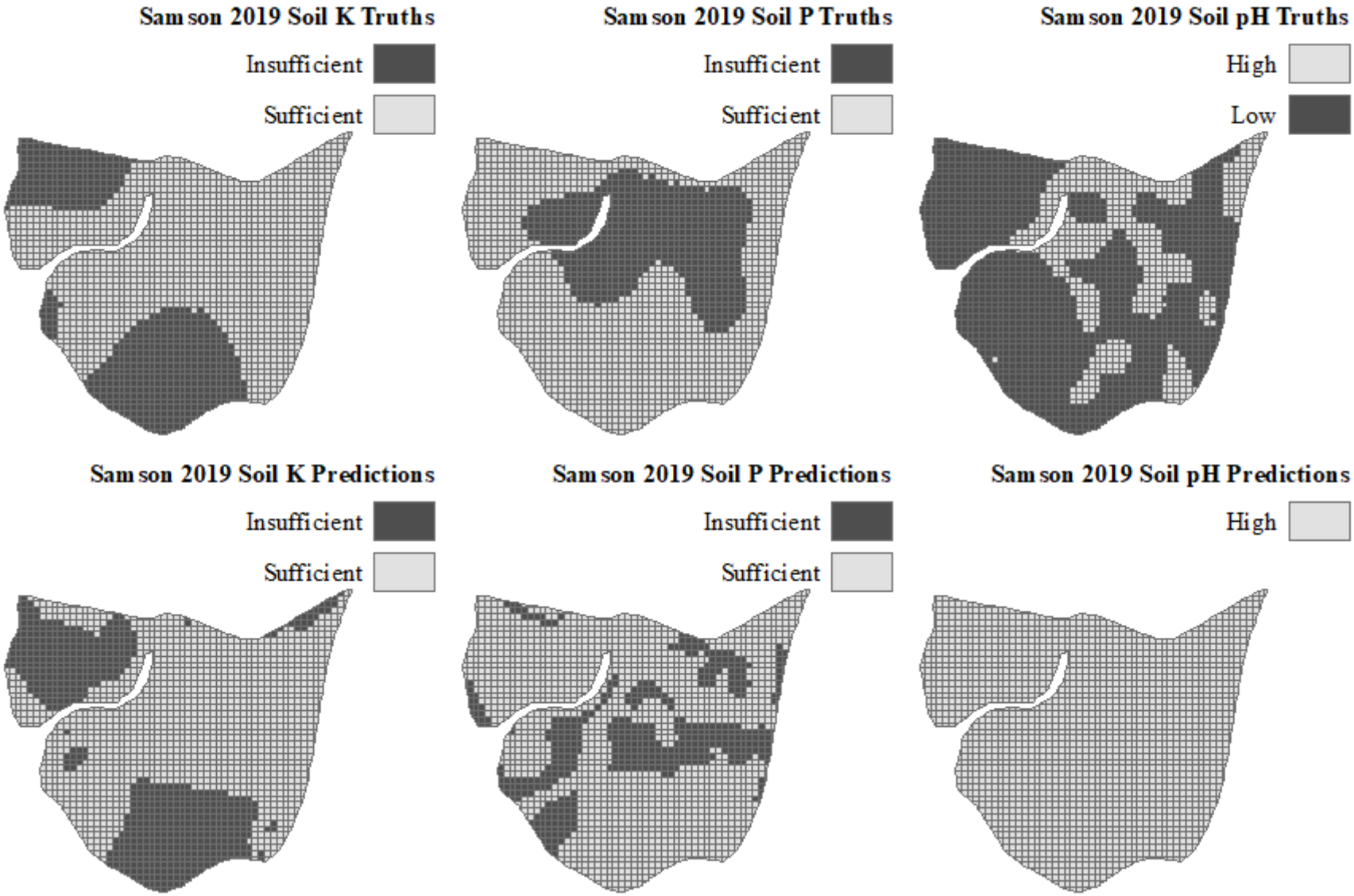
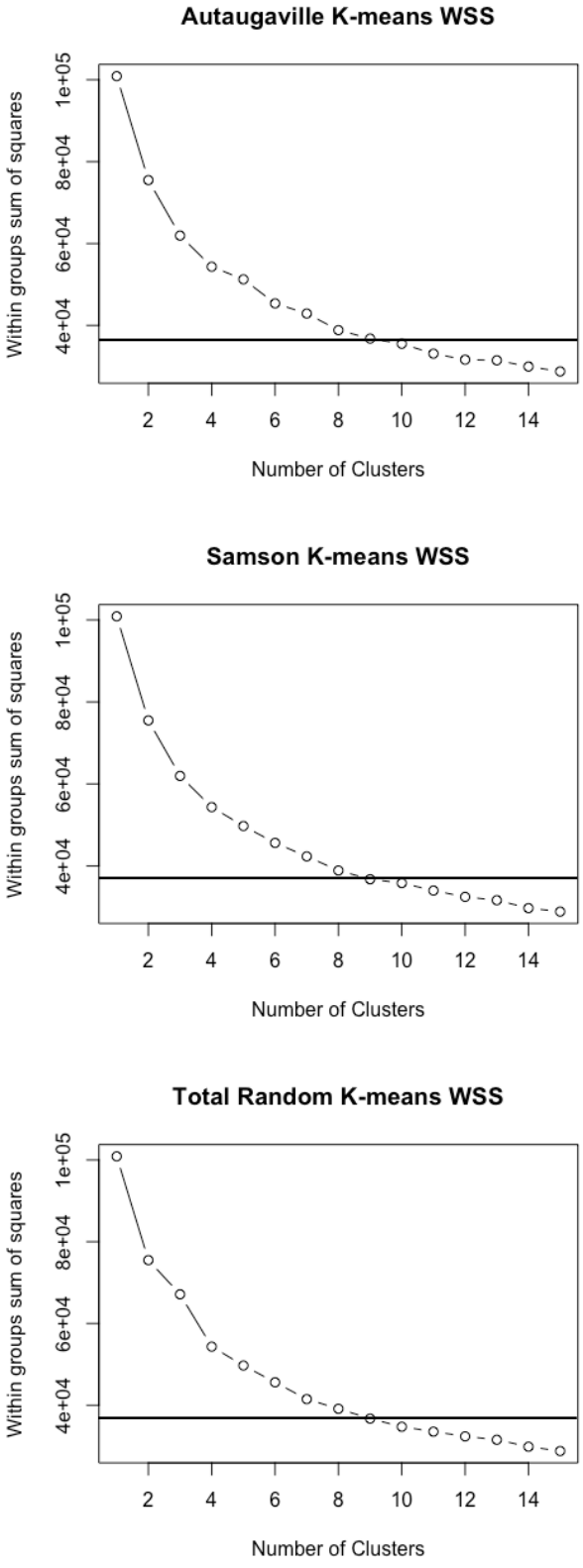


Figure 22. K-means elbow plots to determine cluster quantity



LITERATURE CITED

Agarwal, A., T. El-Ghazawi, H. El-Askary, and J. Le-Moigne. 2007. Efficient Hierarchical-PCA Dimension Reduction for Hyperspectral Imagery. 2007 IEEE International Symposium on Signal Processing and Information Technology. p. 353–356

Ben-Dor, E., and A. Banin. 1995. Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties. Soil Science Society of America 59: 364–372. doi: 10.2136/sssaj1995.03615995005900020014x.

Boettinger, J.L., R.D. Ramsey, J.M. Bodily, N.J. Cole, S. Kienast-Brown, et al. 2008. Landsat Spectral Data for Digital Soil Mapping. In: Hartemink, A.E., McBratney, A., and Mendonça-Santos, M. de L., editors, Digital Soil Mapping with Limited Data. Springer Netherlands, Dordrecht. p. 193–202

Bray, R.H., and L.T. Kurtz. 1945. Determination of Total Organic and Available Forms of Phosphorus in Soils. Soil Science 59(1): 39.

Demšar, U., P. Harris, C. Brunsdon, A. Fotheringham, and S. Mcloone. 2013. Principal Component Analysis on Spatial Data: An Overview. Annals of the Association of American Geographers 103. doi: 10.1080/00045608.2012.689236.

Ferguson, R.B., and G.W. Hergert. 2000. Soil Sampling for Precision Agriculture. University of Nebraska Extension EC 00-154: 4.

Fox, G.A., and R. Metla. 2005. Soil Property Analysis using Principal Components Analysis, Soil Line, and Regression Models. Soil Science Society of America Journal 69(6): 1782–1788. doi: 10.2136/sssaj2004.0362.

Godwin, R.J., and P.C.H. Miller. 2003. A Review of the Technologies for Mapping Within-field Variability. Biosystems Engineering 84(4): 393–407. doi: 10.1016/S1537-5110(02)00283-0.

Kerry, R., and M.A. Oliver. 2008. Determining nugget:sill ratios of standardized variograms from aerial photographs to krige sparse soil data. Precision Agriculture 9(1–2): 33–56. doi: 10.1007/s11119-008-9058-0.

Ladoni, M., H.A. Bahrami, S.K. Alavipanah, and A.A. Norouzi. 2010. Estimating soil organic carbon from soil reflectance: a review. Precision Agric 11(1): 82–99. doi: 10.1007/s11119-009-9123-3.

López-Granados, F., M. Jurado-Expósito, S. Atenciano, A. García-Ferrer, M.S. de la Orden, et al. 2002. Spatial variability of agricultural soil parameters in southern Spain. Plant and Soil 246(1): 97–105. doi: 10.1023/A:1021568415380.

Mehlich, A. 1953. Determination of P, Ca, Mg, K, Na, and NH4. North Carolina Soil Test Division.

Mehlich, A. 1984. Mehlich 3 soil test extractant: A modification of Mehlich 2 extractant. Communications in Soil Science and Plant Analysis 15(12): 1409–1416. doi: 10.1080/00103628409367568.

Qiao, Y., and S. Zhang. 2012. Near-Infrared Spectroscopy Technology for Soil Nutrients Detection Based on LS-SVM. In: Li, D. and Chen, Y., editors, Computer and Computing Technologies in Agriculture V. Springer Berlin Heidelberg. p. 325–335

Uno, Y., S. Prasher, R. Patel, I. Strachan, E. Pattey, et al. 2005. Development of field-scale soil organic matter content estimation models in Eastern Canada using airborne hyperspectral imagery. Canadian Biosystems Engineering / Le Genie des biosystems au Canada 47: 1.9-1.14.

Xin-Zhong, W., L. Guo-Shun, H. Hong-Chao, W. Zhen-Hai, L. Qing-Hua, et al. 2009. Determination of management zones for a tobacco field based on soil fertility. Computers and Electronics in Agriculture 65(2): 168–175. doi: 10.1016/j.compag.2008.08.008.