Imputing plant community classification from associated forest inventory and physiographic data in Minnesota, USA

David C. Wilson^a and Alan R. Ek^a

^a Department of Forest Resources, University of Minnesota,
115 Green Hall, 1530 Cleveland Ave. N., St. Paul, MN 55108, USA, wils0602@umn.edu, ae@umn.edu.

Abstract

We use conditions defining growing space (e.g., sum of all relevant biotic and abiotic factors) to systematically associate forested stands with likely native plant community (NPC) classifications. We develop and test imputation methodology to link ecological information with forest inventory records, i.e., to individual stands. These efforts will reduce the cost of obtaining needed ecological data with the detail and site specificity required for regional landscape-scale management. We employ associative techniques similar to those used by the Minnesota Department of Natural Resources native plant sampling program to group sites sharing multiple conditions and plant species. However, the focus here is on linking plant community observations with conditions observed during routine forest stand inventory efforts. Unique associations of herbaceous plants are thought to occur in conjunction with different sets of trees on sites with appropriate soils, physiography, moisture, and disturbance histories. Following this hypothesis, NPC assignments based on field observation of herbaceous and woody plants are used as a source of training data to identify and categorize stands already characterized by routine forest inventory efforts. Additional soils, climate, and physiographic data have also been associated with NPC and forest inventory observations to provide information about shared characteristics influencing vegetation on a site. Imputed classifications provide 4 levels of information about ecological conditions on a stand; ecological system, floristic region, and moisture and nutrient availability. Results suggest that with 13 ecological systems, we might expect an average of $70\% \pm 10\%$ of imputations to be correct at the system level. Imputation of floristic region, moisture index, and nutrient index ratings resulted in roughly 90% accuracy for each component. The compounding of small error rates in components of the NPC class code leads to a $47\% \pm 10\%$ rate of success over failure in imputation of the class a stand might later be assigned based on relev data. Relative correspondence of imputed classes to those assigned through observation (e.g., similarity to the true classification; precision) is roughly 87%. Essentially, this means that even when incorrect, imputed classes are very close to those assigned based on observation of physical and vegetative associations. This level of success will help bio-regional landscape management efforts through savings in NPC mapping and classification efforts, and the ecological understanding provided.

Keywords: forest inventory, native plant community, classification, imputation, physiography, maximum likelihood.

1. Introduction

A driving force behind this research is an interest in using detailed ecological information in the process of developing and selecting appropriate management options for forested stands and ecosystems. The desire to make use of detailed ecological information pre-supposes a knowledge and understanding of the distribution, and successional state of various native plant communities (NPCs) composing the local and bio-regional ecosystems. NPC classifications are based on 4 class code components (Figure 1) defining the ecological system, floristic region, moisture and nutrient conditions occurring at a site. Unfortunately, such information is only rarely available with the level of detail, coverage, and site specificity needed to make timely decisions for management (e.g., harvest scheduling, wildlife habitat enhancements, water quality protection, maintenance of biodiversity, and other values) at multiple scales. This research seeks to bridge this gap by leveraging what we do know about NPC distribution and landscape associations to assign appropriate NPC classifications to individual forest stands.

Native plant community associations¹ develop on a landscape scale over many years (e.g., decades to millennia). Different sites will foster the development of different NPC associations based on physiographic and climatic conditions, the propagule sources prevailing locally, and disturbance history. The time of year a disturbance occurs, its intensity, and the time between disturbances all influence plant community development. Changes in these factors produces great variation in the structure and composition of resulting communities (MNDNR 2003). Terrain, soil characteristics, and prevailing climate are also related to the frequency, intensity, and timing of fires, drought, windstorms, ice storms, and extreme snowfall events. These relationships are thus suggestive of likely predictor variables related to NPC classification.

Because plant community development is strongly influenced by the combined growing space (e.g., sum of all biotic and abiotic factors affecting plant development) and disturbance history of a site, it is possible to use existing vegetation and abiotic conditions to identify similarities among sites on the landscape (Aaseng et al. 2011, MNDNR 2013). It is important, however, to note that observed patterns may be related to either physical site characteristics, disturbance history, or both. As discussed above, terrain and climate largely control the frequency and intensity of fires, windstorms, and precipitation, so these factors, combined with vegetative observations, can be used as a surrogate where disturbance history is unknown. We can then group sites according to their physical qualities and the plant associations best able to make use of the growing space represented by that site. The extent to which existing vegetation on a site has been shaped by the interplay of dispersal, competition, succession, and periodic disturbance determines the likely success of using the methodology developed here to identify the same NPC class(es) that would be selected using detailed observation of plant associations.

¹ Native plant community classes are defined as units of vegetation that generally have uniform soil texture, soil moisture, soil nutrients, topography, and disturbance regimes (MNDNR 2003). MNDNR currently recognizes approximately 99 distinct NPC classes, with additional types and subtypes defined for many (MNDNR 2006).

[Figure 1]

Growing space is shaped by physical site characteristics, but also by the effects of inter and intra-specific competition for light, nutrients, moisture, and less obvious needs. For these reasons, dominant tree species, successional state, and the extent to which a site has been captured by tree canopy are relevant to the growing space available for both tree and herbaceous species to become established and persist on a bioregional landscape. Herbaceous plant associations provide additional information helpful to differentiating finer NPC classifications (Figure 1) related to variations in regional to microsite scale conditions (Aaseng et al. 2011, MNDNR 2003, MNDNR 2013).

A geographic information system (GIS) is defined as “a system designed to capture, store, manipulate, analyze, manage, and present spatial or geographic data” (Wikipedia.org). The recent proliferation of remote sensing and “big geographic data”, along with the rapid development of GIS capabilities, has provided the opportunity to conduct large-scale predictive analyses of landscape ecological condition. These kinds of analyses typically use a variety of thematic layers describing relevant physical and biotic characteristics of the landscape to make broad prediction of ecological context. Such geographic layers may include themes like land cover, percent forest canopy cover, precipitation, soil drainage, soil fertility, slope, elevation, ecological context, or other relevant factors. Here, we use these broad characterizations combined with selected data layers and field observations to create finer ecological classifications associated with distinct plant communities present on forested stands, or management units, administered by MNDNR.

Existing methods for extending observations to classify very similar nearby entities (e.g., forest inventory units) by nature rely on nonparametric approaches able to accommodate categorical predictor and response variables. The randomForest approach (Breiman 2001, Liaw and Wiener 2002, Cutler et al. 2007, Packalén et al. 2012), which evaluates a large number of randomly generated classification and regression trees (CARTs) created using different subsets of the observations and predictors, is one such machine learning tool. Withheld, or out of bag, observations (~37% of the data) are used to assess the performance of each CART. Variable importance is assessed through permutation-based mean squared error (MSE) reduction associated with different predictors across the forest of CARTs. Unfortunately, randomForest is limited in the number of categorical predictor levels it can accommodate, and fails to classify entities with incomplete or missing predictor data. Similar tools, including mice (van Buuren and Groothuis-Oudshoorn 2011) and earth (Milborrow 2016), suffer from the same limitations, and also provide incomplete coverage of entities we might wish to classify.

This research explores methodology and performance of more flexible imputation modeling designed to identify most likely NPC class associations for forested stands managed by Minnesota Department of Natural Resources (MNDNR). In doing so, the hypothesis that tree cover and abiotic conditions are highly correlated with the unique associations of plants

identified by field ecologists using the relevé method (Minnesota Department of Natural Resources, 2013) is tested. Hence, performance of the imputation model developed here can also be thought of as a further test of the hypothesis that systematically sampled, detailed vegetation lists are truly necessary for classification of forested plant communities at some desired level of detail. Additionally, the hypothesis that inclusion of a detailed ecological predictor (e.g., land type association; Cleland 1997) in the imputation model will enhance classification accuracy is also tested. Implications of the imputation model for relevé survey and forest inventory design, cost, and timing are also explored.

2. Material and methods

2.1 Data and Pre-Analysis

We begin with a review of available data tracking variables related to forest ecosystem development and growth in Minnesota. This includes soil, landform, temperature, moisture, and nutrient availability, as all are important factors contributing to native plant community development.

2.1.1 Forest Inventory

MNDNR Division of Forestry maintains a Forest Inventory Management database (FIM), used to track stand inventory data for the 202,151 stands (5.5 million acres – 2.2 million hectares, 4.17 million forested acres – 1.7 million forested hectares) administered by the agency. Observations detailing major forest cover type, tree species present and order of importance, stand age, basal area, timber volume, physiographic class, topography, and other site characteristics are recorded in this data set (Minnesota Department of Natural Resources, 1981). Most stands have been sampled in the past 10 years. However, many of the slower growing swamp conifer stands have not been visited in up to 30 years.

2.1.2 Native Plant Community Observations / Ecological Classification System

A total of 16,913 native plant community observations made by MNDNR Division of Ecological and Water Resources and others between 1964 and 2015 (Minnesota Department of Natural Resources, 2015) are employed in this research. The publication *Field Guide to the Native Plant Communities of Minnesota: The Laurentian Mixed Forest Province* (Minnesota Department of Natural Resources, 2003) provides an overview of the classification process. This guide also contains a dichotomous key to help distinguish finer details of NPCs found in that ecological province. The MNDNR field methodology for NPC sampling is further detailed in the *Handbook*

for collecting vegetation plot data in Minnesota: *The Relevé Method 2nd edition* (Minnesota Department of Natural Resources, 2013).

Spatial data mapped under the National Hierarchical Framework of Ecological Units (Cleland et al. 1997) were also used to inform the imputation model. Units corresponding to ecological subsections and land type associations were used to provide information relevant to potential plant community development on regional and local scales respectively.

2.1.3 Soils

The Soil Survey Geographic Data Base (SSURGO) produced by United States Department of Agriculture Natural Resource Conservation Service (Natural Resources Conservation Service, 2016) has not been completed for parts of northern Minnesota. Consequently, an alternative, the Minnesota Soil Atlas (LMIC 2004), was used. The Minnesota Soil Atlas is a geospatial dataset compiled by the University of Minnesota's Department of Soil, Water and Climate in collaboration with Minnesota's Land Management Information Center (LMIC; now MnGeo) from detailed soil surveys published by NRCS between 1969 and 1981. Low resolution spatial data (40 acre – 16.2 hectare mapping unit) included in this mapping effort describe land position, soil drainage, soil texture, soil landscape unit, thickness of the rooting zone, and available water, potassium, phosphorus, and PH across the entire state. Because the intent of this research is to characterize conditions on the scale of a forest stand (10-1,000 acres – 4 - 405 hectares), with wall-to-wall coverage, data layers included in the Minnesota Soil Atlas are examined here. Data layers describing soil drainage (Schaetzl et al. 2009) and productivity (Schaetzl et al. 2012) for the lower 48 states at a 250m resolution were also assessed for use with the imputation modelling.

2.1.4 Climate Data

Additional information describing predominant growing season (e.g., May – September) climate over 30 years (1981 – 2010) was also used to inform the model (PRISM 2016). Climate variables considered included average maximum temperature and total precipitation observed during the period of primary interest.

2.1.5 Geospatial Analysis

Initially, ArcGIS10.2 (ESRI 2013) was used to create a spatial join between FIM stands and NPC relevé observations, thereby linking attributes present in the NPC layer to FIM stands. This process was completed using one-to-many (e.g., one stand may have multiple corresponding NPC relevés) intersect and spatial join functions. Next, the hybrid FIM-NPC spatial data was

associated with additional soils, landform, climate, physiographic, and ecological data. This process used the “Extract Multi Values to Points” tool (available through the Spatial Analyst Tools toolbox) in ArcGIS 10.2, using the centroid of each FIM stand as the target for the new data association. Augmented forest inventory and NPC data were then used to assess variable importance and association of predictor variables with NPC class observations.

2.1.6 Predictor Selection

Initial predictor selection used the package randomForest (Breiman 2001, Liaw and Wiener 2002, Cutler et al. 2007, Packalén et al. 2012) in R (R Core Team 2015) to select variables with the highest relative importance to NPC classification. Both the mean decrease in accuracy and the decrease in the Gini Index of inequality (Gini 1912) were used to assess predictor importance. The mean decrease in accuracy quantifies the loss of predictive power when a variable is dropped from the model. The Gini Index, from the socio-economic literature, is used here as a measure of predictor efficiency in differentiating classes. The dual selection approach enabled by randomForest yielded a different ranking of predictor variables from each method (Table 1).

A surrogate for tree species dominance status was created by ranking each tree species by the proportion of basal area it contributes to the stand. Direct observation of tree crown class (e.g., proportion of stem supporting live branches) can also be used, if available (e.g., for United States Department of Agriculture, Forest Service, Forest Inventory and Analysis – henceforth FIA; Miles 2014), and may provide a more reliable dominance metric (Wilson and Ek 2017). The first through third most dominant tree species by basal area were considered in this analysis. The fourth most dominant tree species (e.g., as ranked by basal area for MNDNR, or classified as understory by FIA) was omitted due to its lesser predictive importance and inconsistent characterization by different forest inventories. The more detailed Ecological Classification System (ECS) land type association (LTA) had too many factor levels (203 LTAs in Minnesota) for inclusion in the randomForest importance analysis, but was assessed separately using methodology described here. Limitation of missForest, the imputation component of randomForest, to categorical variables with 53 levels or fewer, results in the need for this separate assessment of importance. This limitation also required a simplification of the dependent variable (NPC Class) to just the 52 most common classes plus a combined class for all more rare plant communities. Any or all of the predictors ranked among the top 8 by the variable importance analysis, plus LTAs were used in the final imputation to identify likely NPC associations.

[Table 1]

2.2 Calculation / Imputation

Imputation is, essentially, the assignment of likely values or characteristics to entities for which we have incomplete data based on associations made with similar entities for which we do have complete data. Observed characteristics of the entity to be imputed serve to inform the most likely value or characteristic chosen by the imputation method. In this case, we have 16,913 NPC training observations collected between 1964 and 2015 by MNDNR and others (MNDNR 2015). Each of these observations falls within a polygon outlining a management unit, usually a forested stand, under the administration of MNDNR. Relevé locations used in this analysis are typically chosen by the field researcher to be a high quality, representative sample of the vegetation in that stand, and avoid any signs of human-related disturbance (MNDNR 2013).

Initial attempts to use pre-existing imputation packages proved unsatisfactory. Both `missForest` (Stekhoven and Bühlmann 2012), an add-on to the popular `randomForest` package (Breiman 2001, Liaw and Wiener 2002, Cutler et al. 2007, Packalén et al. 2012), and `earth` (Milborrow 2016) were used in R (R Core Team 2015) to perform test classifications. Input records included only those with complete predictor data, so the only variable being imputed was NPC class, rather than a random subset of missing values for both dependent and independent variables. The results from both packages were similar, with 42-48% positive predictive value (e.g., correct NPC class assignment), and roughly half of all missing classes not imputed. The R package “`mice`” (Multivariate Imputation by Chained Equations - van Buuren and Groothuis-Oudshoorn 2011) was also used in a separate attempt to perform the desired classification. Unfortunately, the architecture of the “`mice`” programming was unable to accommodate all of the factor levels (> 65 million possible combinations) associated with the predictor and response variables. Because of the relatively poor performance (e.g., coverage and accuracy) of these pre-existing imputation packages for the complex categorical associations involved here, alternative methodology was developed.

First, the data were stratified into 3 broad land cover types corresponding to forest (12,746 records), non-forest (497 records), and non-stocked (3,178 records) land. An additional 492 records contained incomplete NPC class codes, and were removed from the analysis. Non-stocked stands included grasslands, brush lands, and other land types that could support trees, but currently do not have substantial tree cover. Data from the corresponding strata were then used to inform imputations for stand polygons. Finally, a 3 step process (Figure 2) founded in record linkage theory (Fellegi and Sunter 1969), rough set theory (Pawlak 1982), and maximum likelihood estimation (Fisher 1925) is used to identify the most likely NPC association based on matches between the physical and biotic characteristics of known NPCs and unclassified stands. This process uses characteristics from a focal polygon to define an approximation space, or query, composed of stand records with characteristics (e.g., predictor variables) matching those of the focal stand. This approximation space encompasses a cluster of training records which are indistinguishable, or homogenous, in terms of the antecedents, but may differ in their NPC classification. From a statistical modelling perspective, each antecedent fixes one of the model

predictors at a specified level. Likelihoods associated with different NPC outcomes in the smaller approximation space are then compared to the global likelihood of the same outcome.

[Figure 2]

Maximum likelihood techniques comparing the frequency of observed NPC classes in the approximation space to the null hypothesis of likelihood proportionate to globally observed frequencies are used via the likelihood ratio test (LRT) (Wilks 1938). Fellegi and Sunter (1969) claim that the optimal linkage rule (e.g., the maximum likelihood estimate, or MLE, given the data) is equivalent to the likelihood ratio test. Finally, the most likely NPC class (e.g., the class assignment corresponding to the MLE) among records encompassed by the approximation space is imputed as the most likely NPC association. Because multiple NPC classes were observed on some stands, the second through fifth most likely NPC associations are also kept as additional potential classes present. The process is then repeated for all records in the forest inventory database.

The initial imputation process relied heavily on dynamically generated SQL (Structured Query Language) applied within a logical framework written in Visual Basic for Applications (VBA) (Microsoft 2013). In short, each record in the inventory list was sequentially used as the basis for defining an approximation space in the form of a SQL query. The approximation space corresponds to an exponential family (e.g., multinomial) sub-model defined by predictor variables fixed at specific levels by characteristics of the focal record (e.g., stand). The larger model, used as a reference for the *a priori* distribution of NPC classes, is defined by the union of all relevant predictor levels associated with stands classified via relev observations. The process was then translated to the more analysis oriented R computing and statistical environment (Code Box 1) (R Core Team 2015).

[Code Box 1]

Antecedents defining the approximation space are initially quite stringent, involving up to 9 unique conditions which must be met for inclusion in the cluster of matching training records. If insufficient matching records are returned, conditions are sequentially dropped until a cluster containing sufficient matching records for a Chi-square frequency analysis (e.g., $((\text{obs} - \text{exp})^2 / \text{exp}) \sim \chi^2$; $\text{df} = \# \text{ of parameters fixed for sub-model}$) is found. A lower limit of 5 NPC observations in the approximation space is used here. The use of more conditions (e.g., fixed parameters) in the antecedent for a given sub-model produces a smaller and more precise cluster with less variability in observed NPC classes. The number of parameters fixed for any given sub-model approximation space can then be used in combination with cluster purity (Parmar et al. 2007, Herawan et al. 2010) as a basis for sorting good imputations from more suspect decisions. In this case, the cluster is the set of NPC classes encompassed by the approximation space. Cluster purity is then a measure of classifier homogeneity in the approximation space.

Variables used to define approximation spaces include average summer precipitation, average maximum summer temperature, drainage index, ecological subsection or land type association, forest cover type, primary tree species, basal area, and land position. Predictors were relaxed sequentially in order of importance to allow for greater coverage of the resulting imputation. The number of predictors used in each imputation was recorded as the degrees of freedom associated with the exponential family sub-model for the decision space.

Evaluation of the relative likelihood of different NPC classes included in the set of matching records was performed using LRT methodology with degrees of freedom equal to the number of fixed parameters used in the antecedent for the sub-model approximation space. This test provides information on the likelihood of finding at least the observed number of records matching a given class in a cluster of the observed size. Both p-values and total number of matches in a given class were used to calculate a variable called Evidence:

$$Evidence_i = \frac{Count_i - 1}{\exp(p.value_i) - 1},$$

where i refers to elements in the list of NPC classes matching a given antecedent.

This evidence based approach helps to screen out random noise occurring when a single unusual NPC match for the given antecedent occurs by chance, while still amplifying the signal from unusually frequent classes. This amplification is proportionate to the likelihood ratio associated with the observations of a given class in the approximation space.

The LRT statistic is calculated as:

$$LRT_i = 2 * \ln\left(\frac{L.hat_i}{L.null_i}\right),$$

where $L.hat_i$ is the proportion associated with distinct NPC classes present in the approximation space:

$$L.hat_i = \left(\frac{Count(NPC_i)}{Count(NPC_{all})}\right),$$

and $L.null_i$ is the probability of sampling the given class from the full set of NPC observations:

$$L.null_i = \left(\frac{Total(NPC_i)}{Total(NPC_{all})}\right),$$

To allow analysis of model performance for each of the 4 NPC class code components (Figure 1), individual summarizations of the approximation space were also created for each class code component. Results are then ranked by Evidence and by LRT statistic to determine their position in the cluster of imputed classes for the given antecedent.

Model selection using a modified AIC (Akaike 1973) approach indicates that a model allowing for the cumulation of imputations made with progressively fewer predictors should be preferred for greatest parsimony and coverage. A formulation of AIC similar to that used by Figueiras and Cadarso-Suárez (2001) is employed, with the addition of a penalty for lack of coverage. This method of cumulating imputations bears a striking resemblance to the generalized affine functions described by Geyer (1990), in which the supremum function corresponds to the full set of classifiers (and the likelihoods of those classifiers) imputed using decreasing model dimensionality until the MLE for all objects in the database has been identified. Such a function is described by the definite integral:

$$\int e^h d\lambda = \int_{cl D} e^h d\lambda,$$

where $cl D$ (closure D) is the domain of all log likelihoods (h) associated with potential classifiers (λ). The cumulant function then identifies the MLE (θ), which maximizes:

$$c(\theta) = \int e^{<x,\theta>} d\lambda(x),$$

across predictors (x) defining each face (e.g., hyperplane) of the generalized affine model. The main difference between the current approach and models discussed by Geyer (1990) is that the impMLE analysis keeps the classifier identified using the highest dimension model providing sufficient support for a decision (e.g., ≥ 5 matching records).

2.2.1 Application to a New Population

Finally, bootstrap validation was used in two separate approaches to explore likely model performance when applied to a new population of forest inventory stands. This process examines the possibility that despite efforts to provide a representative sample of NPCs on the landscape, the training data might be missing some portion of total variability present in the population (e.g., the full set of 202,151 stands managed by MNDNR). At the outset of this research, 11,131 complete NPC relevés sampled on forestland prior to 2013 were available. An additional 2,301 complete relevés were observed on forested stands during the 2013 - 2015 field seasons. Complete relevés are considered to be those classifying a site with a full four part class code (e.g., over 400 incomplete classifications of forested stands were not used here).

The first validation procedure assumes that NPC samples available at the onset of the project are independent from relevés collected during the 2013 – 2015 period. Following this assumption, repeated samples of 100 records each were taken from the newer data for use in testing the model. Each small sample of more recently observed stands was then imputed using the older set of observations to inform the model. Imputations were then compared with observations to

repeatedly assess performance of the model. Over 1,000 repetitions, this process explores the likely variability in model performance when applied to a new sample of 100 FIM stands.

The second bootstrap process uses all of the relev data to inform the model, but enforces an out-of-bag association. This out-of-bag process screens observations corresponding to stands in the focal block of test records from the training data used for that iteration of the bootstrap. Test data sets were repeatedly constructed by resampling the full set of observed stands, with replacement.

2.2.2 Accuracy Assessment

Performance of the imputation model was assessed using two criteria derived from the ratio of correct imputations to total imputations for each NPC class code component. The measure **Accuracy** is calculated as $Accuracy = System * Region * Moisture * Nutrient$. Accuracy for **System** and **Region** are Bernoulli distributed (e.g., 0 or 1), while **Moisture** and **Nutrient** reflect the relative distance of the imputed indicator from observations on a 0 to 1 scale, where 1 indicates perfect agreement. The measure **Precision** is calculated as $Precision = (System + Region + Moisture + Nutrient)/4$. **Precision** provides a measure of how similar imputed NPC codes were to the observed NPC class.

3. Results

Results of testing on the new data for forested stands (Table 2a.) were similar to the out of bag bootstrap analysis (Table 2b.). When considering combined class code components of the most likely imputation, mean absolute **Accuracy** was estimated at 53.44% with an approximate 95% confidence interval of $\pm 10\%$ for the split data (e.g., pre-2013 vs. 2013+). The out of bag bootstrap suggests that roughly $55.12\% \pm 10\%$ accuracy can be expected when applying the model to a truly independent population. Results for **Precision** (mean = 0.867 ± 0.03) show that imputed NPC class codes are typically quite close to the codes arrived at through analysis of relev data, with errors concentrated in imputation of ecological system.

Bootstrapped imputation performance measures were normally distributed for 1,000 sets of 100 stands resampled from the set of observed stands. For the measures with a range of 0 to 1, this distribution could be described as $N(\pi, \sigma)$, where π is the mean value statistic for the set of 1,000 small samples evaluated here, and σ , the standard deviation of the statistic, is approximated by $0.5 * (UPR.95 - LWR.95)/1.96$. See Figure 3 for examples of two of these model performance distributions. In short, this means that we can provide an estimate of expected imputation performance with known bounds of certainty for both whole NPC codes and imputed code components when the model is applied to a new population.

[Table 2]

[Figure 3]

An assessment of imputation results compared against those from missForest is presented in Table 3. To summarize, when degrees of freedom associated with imputations are used to stratify the results, we can compare performance metrics against those provided by missForest. Overall, the proportion of false classifications provided by missForest for the test data is approximately 52% with 55% coverage (e.g., proportion of missing classifications imputed). A comparison with the current maximum likelihood estimator imputation (impMLE) uses between 2 and 7 predictors to classify 63-100% of missing values, with 48-54% false classifications. Hence, the methodology developed here is capable of using more detailed categorical predictors than can be accommodated by randomForest/missForest, provides greater coverage for the same level of accuracy, and can attain higher levels of accuracy by limiting results to those decisions made using more of the available predictors. Model selection using a modified AIC (Akaike 1973) approach indicates that cumulating results across submodels with decreasing complexity, should be preferred for greatest parsimony and coverage. Thus, impMLE methodology, including the least restrictive submodel with only growing season temperature and precipitation (plus the condition of being forest), are preferred.

[Table 3]

Use of impMLE methodology also allows us to compare the relative contribution to predictive capacity of land type association (LTA) and physiographic class (PHYS) (Table 3). The complex information contained in the LTA predictor does provide additional accuracy in cases where sufficient training data corresponding to each level of the predictor exists. However, inclusion of either PHYS or LTA in the model reduces coverage by masking potential matches with training data differing only by these criteria.

Additional performance measures can be calculated based on the confusion matrix (Table 4) for imputations and observations. However, these metrics provide little additional insight to the performance of this imputation model. Thus, they are noted here as background and possible options for future problems of this type.

[Table 4]

4. Discussion

Physiography is widely described as an important determinant of ecological development (Bowman 1911, Clement 1916, Parker and Weaver 1989). This may be in part because the concept of physiography includes characteristics of soils, landform, and hydrology. Results reported here indicate that the combination of physiography (e.g., land form and drainage) and

climate (e.g., precipitation and temperature), while not perfect, tells us a lot about potential growth factors and, therefore, likely plant community composition. Indeed, it is the joint dependence of trees and herbaceous species on the unique growing space provided by a site which enables the imputation process described herein to function successfully. This is made clear by the closeness of imputed moisture and nutrient indices to the observed values.

Interestingly, the MNDNR physiographic variable performed poorly relative to the separate components of land position and drainage index, and as compared against LTA. This is likely due to the environmental simplification inherent in a variable with only 5 levels (e.g., too coarse), as well as the fact that MNDNR methodology for observation of physiography (MNDNR 1981) includes many observations during snow-on conditions. LTA combines similar information about the physical characteristics of a site observed in association with existing plant communities. This ecologically detailed (perhaps too detailed) predictor provides accuracy improvements over the MNDNR physiographic variable, but substantially reduces coverage for similar model complexity.

Hence, we conclude that inclusion of categorical predictors with too much, or too little, detail can produce a less powerful or less generalizable model due to either overfitting (LTA) or relatively poor correspondence of the predictor to the dependent variable observed for stands in different ecological contexts (PHYS). Instead, more basic physical predictors of environmental and ecological characteristics such as drainage, precipitation, temperature, slope, and flow accumulation may better serve to quantify variability in physiographic conditions and ecological context across NPCs. This observation is made clear by the initial predictor importance analysis where the 5 level physiographic class variable barely made the cut, and LTA presented randomForest with too many levels to assess. Instead, the variable importance analysis identified growing season temperature and precipitation (both ancillary layers describing basic physical conditions) as the most important predictors of NPC class. Future iterations of this work may benefit from testing and inclusion of additional GIS predictors available from elevation models, LiDAR, classified satellite imagery, and other sources.

Tree cover is also indicative of the broader plant associations likely to occupy a site. However, tree cover alone does not inform us sufficiently as to the combination of herbaceous species present on the landscape and able to disperse to and dominate a site following disturbance. Tree species present on a site are indicative of conditions during stand initiation. However, for older stands there is an opportunity for a decoupling of historic site conditions from more recent climatic and biotic influences. This shortcoming of the methodology used here is apparent in the relatively lower accuracy found for the imputation of ecological system (roughly 70% correct imputations). This component of the NPC classification is most dependent on the association of herbaceous and woody plants found to actually be occupying a site. We can therefore conclude that inclusion of additional vegetative data (e.g., observations of herbaceous species and/or additional tree dominance data as provided by FIA) in the imputation model may improve results by reducing error in the imputation of ecological system.

Compounding the errors shown in Table 2b for ecological system (70.7% Bernoulli accuracy) with the relatively smaller errors for floristic region (89.3% accuracy), moisture (93.3% accuracy), and nutrients (93.5% accuracy) results in the overall **Accuracy** rate of 55.1% for imputed class codes. These observations suggest that if the presence of certain herbaceous indicator species associated with the broad ecological systems were routinely recorded during forest inventory efforts (or added to the FIM data from associated Ecological Classification System sheets), it might be possible to improve the weakest link of the imputation process. Indeed, it is this very information collected on relevés which informs the conventional classification scheme used by the MNDNR. Adding these observations to the digital FIM database, and ensuring that relevé data are retained in the database after stands are harvested, could help to better differentiate among ecological systems and speed improvements in our knowledge.

While the interpretation of vegetation survey data systematically collected on relevés is the preferred method to positively identify NPC classes present, substantial information is gained through the imputation process. Variable accuracy/precision of imputed NPC classes, suggests that the intensity of observation needed for some stands could be adjusted in order to simply identify the correct NPC from the short list of likely classes resulting from imputation. Knowing the relative accuracy of imputations for different classes allows for a stratification of sampling intensity based on known patterns of error. Also, differentiating among 2-3, or even 5, similar classes should, on average, be easier and less expensive than the sampling needed to differentiate among the full set of classes. An observer familiar with various ecological indicator species prevalent in the region would likely be able to make the distinction with just a few observations (e.g., not full relevés) taken on or between concurrently observed forest inventory plots.

Taken together, these results could produce several benefits to the NPC data collection process and help with incorporating ecological detail into management planning. While it will likely take another 10 years, under current plans, to collect such data on all stands managed by MNDNR Forestry, imputation gives data now, or at least soon for many stands. Further, we can expect gains in final survey accuracy from improved stratification of the field sampling. For example, the MN Biological Survey (a program of the MNDNR Division of Ecological Resources tasked with sampling and documenting Minnesota's diverse biological resources) can focus field work on imputed classes with lower expected accuracy and cluster purity (e.g., where additional observations are most needed). Gathering additional observations from the set of inadequately sampled stands (e.g., those exhibiting variability outside of the conditions observed thus far) will improve the representativeness of our known sample, and will likely improve the accuracy of subsequent imputations.

Further, the intersection of polygons for which we have associated NPC observations with permanent sample plots used in other inventory efforts (e.g., FIA) allows for the extension of the imputation process to these other entities (Wilson and Ek 2017). Similarly, the use of a common forest inventory data format by many state and county agencies in Minnesota potentially allows

for the extension of the process to 2.8 million acres (1.13 million hectares) of County managed land. However, the predominantly agricultural history of tax forfeit County lands in Minnesota may complicate the associative process. These former agricultural lands are, by definition, not representative of relevé locations dominated by high-quality native vegetation in an un-disturbed condition. Consequently, existing relevés may not serve as appropriate training data for classification of novel plant communities found on these lands. A modified relevé methodology focusing observations on the plant associations coming to dominate these disturbed sites might help. However, the plant associations found would likely differ substantially from those dominating less anthropogenically disturbed ecosystems.

5. Conclusion

By making reasonably accurate ecological data broadly available in the form of forested stand polygons and inventory units with associated NPC assignments, the imputation model presented here will greatly enhance our ability to assess bioregional and landscape scale management priorities across Minnesota. Because NPC classifications are mapped to the individual management and sampling units, this information will enable further analysis of landscape composition, successional state, and potential ecological trajectories given various management options and climate scenarios. Such information has rarely been available with the spatial coverage, ecological detail, and site specificity enabled by the research presented here.

Herbaceous plant observations do allow for finer and more accurate distinction of native plant associations. Lesser performance of the imputation model specifically for the determination of ecological system clearly indicates the need for additional vegetative inputs to the model. Additional programming effort could allow for the incorporation of shrub and ground cover data sporadically present in the FIM dataset. However, these observations are infrequent and do not replace the detailed lists of herbaceous species gathered through relevés. Ecological Classification System (ECS) sheets completed by field personnel could provide a source for this additional data, if they were systematically digitized and associated with the FIM database. Ideally, the presence of any unique indicators of ecological system would be noted in the ECS sheets, and be available for association with appropriate FIM records.

Despite relatively low absolute accuracy (e.g., was the observed NPC class imputed?), the patterns of error in the imputations are encouraging. More common and more thoroughly sampled NPCs were also more frequently and more accurately imputed (Wilson and Ek *In Review*). Imputation results for rare classes are dramatically more variable in their accuracy. However, even when wrong, imputations were often for a functionally and ecologically similar NPC. This similarity is captured in the difference between the *Exact* and *Accuracy* measures of performance. Imputations for moisture and nutrient indices were typically quite close to the observed values, thereby driving similarities in imputed vs. observed ecological systems (e.g., floodplain forest vs. wet forest, or fire dependent vs. dry mesic hardwoods). Essentially, the

imputation model presented here does arrive at a cluster of NPCs closely related to the observed class, while frequently imputing the observed class as the most likely. The *Precision* metric suggests roughly 14% deviation of information carried in the MLE NPC from information carried in the observed NPC code for the test imputation performed here.

While providing greater accuracy than existing imputation models able to accommodate complex (e.g., a combination of categorical and continuous) predictor data, the model presented here also provides increased flexibility compared with previously available software. This flexibility can be seen in terms of the number of predictors and predictor levels the model is able to use, as well as in how the model adjusts its complexity to provide classifications for nearly all forest inventory units. These features produce a robust model able to identify maximum likelihood classifiers relevant to a broad range of biological and categorical data questions.

In conclusion, imputation results can be used to inform management now. Even while the rate of exact correspondence in imputed vs. observed NPC class is around 50%, the closeness of imputations to observations by several ecological measures renders the results useful for regional and landscape-scale planning. Additional information provided by the imputation clusters can be used to simplify and accelerate the classification process. Error analysis, explored further by Wilson and Ek (*In Review*), highlights groups of imputations likely to be incorrect, thereby identifying inventory units exhibiting non-sampled variation. These groups of stands will, when sampled, provide key data needed to further improve reliability of the imputation model and increase knowledge faster.

Acknowledgements

Funding for this research was provided by the Minnesota Interagency Information Cooperative (IIC) through the Minnesota Department of Natural Resources Division of Forestry, the Minnesota Environmental and Natural Resources Trust Fund as recommended by the Legislative Citizens Commission on Minnesota Resources (LCCMR), the University of Minnesota Department of Forest Resources and the McIntire-Stennis Cooperative Forest Research Program administered through the USDA. The IIC and its member agencies also contributed substantially to the motivation and data needed to complete this research. Thanks also to John Almendinger, MNDNR, who has contributed substantially to the collection, analysis, and classification of relev data for Minnesota's native plant communities. This work would not have been possible without his expertise and participation. The contributions of John Almendinger, Norman Aaseng, and many others working with the Division of Forestry, Minnesota County Biological Survey, Ecological Land Classification, and Natural Heritage and Nongame Research Programs of the Minnesota Department of Natural Resources have provided much of the data, environmental monitoring, and analysis supporting the current research. Further editorial review by Lee Frelich, Matthew Russell, John Fieberg, and two anonymous reviewers have helped to clarify the

573 methods explored here, and their implications for bio-regional landscape monitoring and
574 management in Minnesota and beyond.

575

576

577 Citations

578 Aaseng, N.E., Almendinger, J.C., Dana, R.P., Hanson, D.S., Lee, M.D., Rowe, E.R., Rusterholz, K.A. and
579 D.S. Wovcha., 2011. Minnesota's native plant community classification: A statewide classification of
580 terrestrial and wetland vegetation based on numerical analysis of plot data. Biological Report No. 108.
581 Minnesota County Biological Survey, Ecological Land Classification Program, and Natural Heritage and
582 Nongame Research Program. St. Paul: Minnesota Department of Natural Resources.

583

584 Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. Pages 267–
585 281 in B. N. Petrov and S. Caski, editors. Proceedings of the Second International Symposium on
586 Information Theory. Akademiai Kiado, Budapest, Hungary.

587

588 Bowman, I., 1911. Forest Physiography: Physiography of the United States and Principles of Soils in
589 Relation to Forestry. New York: John Wiley and Sons, Inc.

590

591 Breiman, L., 2001. Random forests. Machine Learning 45: 5–32.

592

593 Cleland, D.T., Avers, P.E., McNab, W.H., Jensen, M.E., Bailey, R.G., King, T., and W.E. Russell, 1997.
594 National Hierarchical Framework of Ecological Units. Published in, Boyce, M. S.; Haney, A., ed. 1997.
595 Ecosystem Management Applications for Sustainable Forest and Wildlife Resources. Yale University
596 Press, New Haven, CT. pp. 181-200.

597

598 Clements, F.E., 1916. Plant succession: an analysis of the development of vegetation. Carnegie Institution
599 of Washington.

600 Cutler, D.R., T.C. Edwards, K.H. Beard, A. Cutler, K.T Hess, J.C. Gibson and J.J. Lawler, 2007. Random
601 forests for classification in ecology. Ecology 88: 2783-2792.

602 ESRI, 2013. ArcGIS Desktop: Release 10.2. Redlands, CA: Environmental Systems Research Institute.

603 Fellegi, I.P. and A.B. Sunter, 1969. A theory for record linkage. Journal of the American Statistical
604 Association 64: 1183-1210.

605

606 Figueiras, A. and C. Cadarso-Suárez, 2001. Application of Nonparametric Models for Calculating Odds
607 Ratios and Their Confidence Intervals for Continuous Exposures. Am. J. Epidemiology 154: 264-275.
608 doi:10.1093/aje/154.3.264

609

610 Fisher, R. A., 1925. Theory of Statistical Estimation. Mathematical Proceedings of the Cambridge
611 Philosophical Society 22: 700-725, doi:10.1017/S0305004100009580.

612

613 Geyer, C.J., 1990. Likelihood and Exponential Families. PhD Thesis. University of Washington.

614 Gini C. 1912. Variabilità e mutabilità: Contributo allo studio delle distribuzioni e delle relazioni
615 statistiche. Studi Economico-giuridici della Regia Facoltà Giurisprudenza 3: 3–159.

616 Herawan, T., R. Ghazali, I.T.R. Yanto, and M.M. Deris, 2010. Rough set approach for categorical data
617 clustering. Int. J. Database Theory Appl. 3: 33–52.

618
619 Land Management Information Center (LMIC), 2004. Minnesota Soil Atlas. Land Management
620 Information Center. 300 Centennial Building; 658 Cedar Street, St. Paul, Minnesota 55155.
621 <http://www.mngeo.state.mn.us/chouse/soil.html>.

622
623 Liaw, A. and M. Wiener, 2002. Classification and Regression by randomForest. R News 2, 18-22.

624
625 Matthews, B. W., 1975. Comparison of the predicted and observed secondary structure of T4 phage
626 lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure 405: 442–451.
627 doi:10.1016/0005-2795(75)90109-9.

628
629 Microsoft, 2013. Office VBA language reference. [https://msdn.microsoft.com/en-](https://msdn.microsoft.com/en-us/library/office/gg264383.aspx)
630 [us/library/office/gg264383.aspx](https://msdn.microsoft.com/en-us/library/office/gg264383.aspx) [Accessed: February 29th, 2016].

631
632 Miles, P.D., 2014. EVALIDator-PC (Version 1.6.00.00) [Software]. Newtown Square, PA: U.S.
633 Department of Agriculture, Forest Service, Northern Research Station. [http://apps.fs.fed.us/fiadb-](http://apps.fs.fed.us/fiadb-downloads/datamart.html)
634 [downloads/datamart.html](http://apps.fs.fed.us/fiadb-downloads/datamart.html)

635
636 Milborrow, S., 2016. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's
637 Fortran utilities with Thomas Lumley's leaps wrapper. (2016). earth: Multivariate Adaptive Regression
638 Splines. R package version 4.4.4. <https://CRAN.R-project.org/package=earth>.

639
640 Minnesota Department of Natural Resources, 1981. Forest survey manual, phase II: intensive inventory.
641 Division of Forestry, MNDNR, St. Paul, MN.

642
643 Minnesota Department of Natural Resources, 2003. Field Guide to the Native Plant Communities of
644 Minnesota: The Laurentian Mixed Forest Province. Ecological Land Classification Program, Minnesota
645 County Biological Survey, and Natural Heritage and Nongame Research Program. MNDNR, St. Paul,
646 MN.

647
648 Minnesota Department of Natural Resources, 2006. Tomorrow's Habitat for the Wild and Rare: An
649 Action Plan for Minnesota Wildlife, Comprehensive Wildlife Conservation Strategy. Division of
650 Ecological Services, MNDNR, St. Paul, MN.

651
652 Minnesota Department of Natural Resources, 2013. A handbook for collecting vegetation plot data in
653 Minnesota: The relevé method. 2nd ed. Minnesota Biological Survey, Minnesota Natural Heritage and
654 Nongame Research Program, and Ecological Land Classification Program. Biological Report 92.
655 MNDNR, St. Paul, MN.

656
657 Minnesota Department of Natural Resources, 2015. Natural Heritage Information System Relevé
658 Database. Division of Ecological and Water Resources, Minnesota Department of Natural Resources, St.
659 Paul, MN, USA. (vegetation plot database; accessed 12/16/2015).

660
661 Natural Resources Conservation Service, 2016. United States Department of Agriculture. Web Soil
662 Survey. Available online at <http://websoilsurvey.nrcs.usda.gov/>. Accessed [02/11/2016].

- Packalén, P., H. Temesgen, and M. Maltamo, 2012. Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. *Canadian Journal of Remote Sensing* 38: 557-569.
- Parker, G.R. and G.T. Weaver, 1989. *Ecological Principles: Climate, Physiography, Soil, and Vegetation*. In: Clark, F. Bryan, tech. ed.; Hutchinson, Jay G., ed. *Central Hardwood Notes*. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Forest Experiment Station: Note 2.01.
- Parmar, D., T. Wu, and J. Blackhurst, 2007. MMR: An algorithm for clustering categorical data using Rough Set Theory, *Data & Knowledge Engineering* 63: 879-893.
<http://dx.doi.org/10.1016/j.datak.2007.05.005>.
- Pawlak, Z., 1982. Rough sets. *International Journal of Parallel Programming* 11: 341–356, doi:10.1007/BF01001956
- PRISM Climate Group, Oregon State University, <http://prism.oregonstate.edu>, created 9 March, 2016.
- R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- W. M. Rand, 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66: 846–850, doi:[10.2307/2284239](https://doi.org/10.2307/2284239). JSTOR [2284239](https://www.jstor.org/stable/2284239).
- Schaetzl, R.J., F.J. Krist Jr., K.E. Stanley, and C.M. Hupy, 2009. The Natural Soil Drainage Index: An Ordinal Estimate of Long Term, Soil Wetness. *Physical Geography* 30: 383-409.
- Schaetzl, R.J., F.J. Krist Jr., and B.A. Miller, 2012. A Taxonomically Based, Ordinal Estimate of Soil Productivity for Landscape-Scale Analyses. *Soil Science* 177: 288-299.
- Stekhoven, D.J. and P. Bühlmann. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28: 112-118. doi: 10.1093/bioinformatics/btr597.
- van Buuren, S. and K. Groothuis-Oudshoorn, 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45: 1-67, <http://dx.doi.org/10.18637/jss.v045.i03>
- van Rijsbergen, C. J., 1979. *Information Retrieval*, London, GB; Boston, MA: Butterworth, 2nd Edition, ISBN [0-408-70929-4](https://www.isbn-international.org/product/0-408-70929-4)
- Wilks, S. S., 1938. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics* 9, 60–62, doi:[10.1214/aoms/1177732360](https://doi.org/10.1214/aoms/1177732360).
- Wilson, D. C. and A.R. Ek, 2017 (*In Press*). Imputing plant community classifications for forest inventory plots. University of Minnesota, Department of Forest Resources. St. Paul, MN 55108. Accepted April 6th, 2017: *Ecological Indicators*.

Tables and Figures

Figure 1. Native plant community class code example.

The system employed for constructing NPC class codes uses four distinct components:
 1.) Ecological system, 2.) Floristic region,
 3.) Moisture regime, and 4.) Nutrient regime.
 Moisture and nutrients are ranked from low (0) to high (9).

Table 1. Variable importance for assignment of NPC class as assessed using randomForest.

Predictors considered in the preliminary analysis of variable importance to NPC classification are ranked by relative influence on success of differentiating among classes. The effect of predictors on success of differentiation is quantified using two methods in randomForest (Breiman 2001, Liaw and Wiener 2002, Cutler et al. 2007, Packalén et al. 2012). First, the mean decrease in classification accuracy observed by removing a variable from the set of predictors is assessed. Second, the mean decrease in the Gini index (Gini 1912) of inequality is calculated as individual predictors are removed from the model. A larger change in the Gini index indicates that a predictor is more efficient in differentiating among clusters of similar classes.

Predictor	Evaluated w/ randomForest	Rank (Accuracy)	Rank (Gini)	Selected	Source
Growing season precipitation	Yes	1	2	Yes	PRISM (1981-2010)
Growing season temperature	Yes	2	1	Yes	PRISM (1981-2010)
Ecological subsection name	Yes	3	4	Yes	Cleland et al. 1997
Land type association	No	NA	NA	Yes	Cleland et al. 1997
Land position	Yes	4	8	Yes	MN Soils Atlas (LMIC 2004)
Drainage index	Yes	5	5	Yes	Schaetzel et al. 2009
Forest cover type	Yes	6	7	Yes	MNDNR (2015) - FIM
Main tree species	Yes	7	6	Yes	MNDNR (2015) - FIM
Basal area	Yes	8	3	Yes	MNDNR (2015) - FIM
Thickness of rooting zone	Yes	9	14	No	MN Soils Atlas (LMIC 2004)
PH	Yes	10	15	No	MN Soils Atlas (LMIC 2004)
Third tree species	Yes	11	11	No	MNDNR (2015) - FIM
Second tree species	Yes	12	10	No	MNDNR (2015) - FIM
Fourth tree species	Yes	13	12	No	MNDNR (2015) - FIM
Productivity index	Yes	14	13	No	Schaetzel et al. 2012
Physiography	Yes	15	16	No	MNDNR (2015) - FIM
Topography	Yes	16	17	No	MNDNR (2015) - FIM
Soil landscape unit	Yes	17	9	No	MN Soils Atlas (LMIC 2004)

724

725

726 **Figure 2. Imputation model process diagram.** First, a set of antecedents is defined by fixing
727 predictor variables at the levels observed for the focal stand record. Next, maximum likelihood
728 methodology (Fisher 1925) is used via the likelihood ratio test (Wilks 1938) to identify the
729 maximum likelihood estimate for NPC classification. Finally, the MLE is imputed as the
730 expected classifier for the focal stand.

731

732

733

```

#Construct a loop to sequentially evaluate each inventory record.

for (i in 1:nrow(impFIM.char)) {

  #Subset training data to select records matching the predictors.
  ##"$Antecedent" refers to one or more matching conditions used.

  NPC.match<-subset(NPCFIM.char, NPCFIM.char$Antecedent ==
impFIM.char$Predictor[i], select=NPC_CLASS)

  #If there are enough matching relevés, enumerate them.

  if(nrow(NPC.match) >= 5){

    summary.NPC<-data.frame(summary(NPC.match))
    colnames(summary.NPC)[1]="Count"
    summary.NPC$Purity<-(summary.NPC["Count"]/sum(summary.NPC$Count))

    #Use Likelihood Ratio Test against hypothesis of proportionate
#probability. A multinomial model is used with df = k fixed
#predictors. Predictors are fixed by matching rules used in the
#antecedent.

    summary.NPC$l.hat<-summary.NPC["Count"]/sum(summary.NPC$Count)
    summary.NPC$l.null<-(NPC[match(row.names(summary.NPC),
NPC$Class),"Prob"]) #NPC is a list of observed class proportions.

    #Calculate LRT statistic, p-value, and Evidence.

    summary.NPC$LRT<- 2*log((summary.NPC$l.hat/summary.NPC$l.null))
    summary.NPC<- summary.NPC[order(-summary.NPC$LRT),]

    #Update stand record to NPC class with largest Evidence and LRT
statistic.

  }

}

```

Code Box 1. Simplified R code for imputation of multivariate class associations

Table 2. Bootstrap performance results for 100,000 imputations. a. Bootstrap imputation performance results¹ for 1,000 sets of 100 imputations testing the split data. b. Distribution of bootstrap imputation performance results for 1,000 sets of 100 out of bag imputations.

a.)

Split data: 1964-2012 vs. 2013-2015								
	Imputed	Exact	System	Region	Moisture	Nutrient	Accuracy	Precision
min	0.8800	0.2778	0.5361	0.7500	0.9021	0.9031	0.3915	0.8054
mean	0.9564	0.4548	0.6795	0.9010	0.9354	0.9322	0.5344	0.8620
max	1.0000	0.6042	0.8229	0.9896	0.9633	0.9594	0.7113	0.9203
upr.95	0.9900	0.5579	0.7742	0.9583	0.9537	0.9495	0.6373	0.8960
lwr.95	0.9100	0.3471	0.5773	0.8333	0.9146	0.9135	0.4343	0.8265

b.)

Out of Bag Imputations								
	Imputed	Exact	System	Region	Moisture	Nutrient	Accuracy	Precision
min	0.8900	0.3402	0.5543	0.7778	0.9032	0.9020	0.4036	0.8131
mean	0.9655	0.4688	0.7073	0.8929	0.9326	0.9348	0.5512	0.8669
max	1.0000	0.6042	0.8542	0.9794	0.9642	0.9579	0.6910	0.9139
upr.95	1.0000	0.5670	0.8000	0.9490	0.9505	0.9516	0.6522	0.9008
lwr.95	0.9200	0.3673	0.6082	0.8265	0.9126	0.9158	0.4540	0.8315

¹ **Imputed** refers to the proportion of unclassified test records receiving an imputed classification. **Exact** refers to imputations for which the first, or most likely, imputation matches one or more observations made on the stand. **System**, **Region**, **Moisture**, and **Nutrient** refer to average accuracy for each component of the NPC class code. **Accuracy** is calculated as **System** * **Region** * **Moisture** * **Nutrient**, and reflects the combined closeness of imputed NPC class components to observations. **Precision** measures how tightly the imputations are clustered around some mean class by averaging the **Accuracy** of each class code component.

a.)

b.)

Figure 3. Distribution of performance measures for bootstrap imputations. a.) Combined accuracy of most likely imputed NPC class code components. *Accuracy* is calculated by combining the means of the Bernoulli distributed (e.g., binary) correctness for ecological system and floristic region with the numeric accuracy (0-1) for moisture and nutrients. N = 1,000 samples, n = 100 imputations. b.) Mean *Precision* of most likely class.

765 **Table 3. Comparison of NPC imputation results for missForest and the current methodology (impMLE) with varying model complexity.** Predictors are
766 numbered according to the following list ranked in order of importance: 1. Growing season precipitation, 2. Growing season temperature, 3. Ecological
767 subsection name, 4. Land position, 5. Drainage index, 6. Forest cover type, 7. Main tree species, 8. Basal area, 9. Land type association, 10. Physiography. Less
768 important predictors were sequentially dropped to allow for greater coverage (# imputed / # of missing values) of the resulting imputation, as indicated by the df
769 (degrees of freedom) column. Degrees of freedom are also summarized as *k* for the cumulation of imputations made using submodels of different complexity.
770 Proportion of false classifications (PFC) is calculated as either 1-*Accuracy* or 1-*Exact* where *Accuracy* and *Exact* are calculated as described in Table 2 and
771 Figure 3. A common model selection approach uses Akaike's Information Criterion (AIC) (Akaike 1973). Here, a formulation of AIC similar to that presented by
772 Figueiras and Cadarso-Suárez (2001) is used (e.g., $AIC = RSS/n + 2 \cdot k \cdot (RSS/(n-k))/n$) with the addition of a penalty for lack of coverage by higher dimension
773 models. The top set of models using missForest is most comparable to impMLE models in the bottom set of rows (shaded white). Criteria for coverage and
774 accuracy of impMLE models with *k* predictors correspond to a summation of those criteria across models with *k* ≥ the count of predictors for that model. In
775 other words, statistics presented for the simplest impMLE models represent a cumulation of results for all submodels with equal or greater complexity.

Tool	Coverage	PFC (1-Accuracy)	PFC (1-Exact)	Predictors	n	df	k	RSS	AIC
missForest	0.4	NA	0.6506	1,2,3,5	12,741	df=4	4	3186.779	1.2545
missForest	0.5	NA	0.548	1,2,3,5,6	12,741	df=5	5	2670.514	1.0521
missForest	0.5714	NA	0.5231	1,2,3,5,6,8	12,741	df=6	6	2425.122	0.9562
missForest	0.625	NA	0.5378	1,2,3,4,5,6,8	12,741	df=7	7	2325.997	0.9178
missForest	0.5556	NA	0.5243	1,2,3,4,5,6,7,8	12,741	df=8	8	2468.441	0.9748
impMLE (LTA)	0.9382	0.5237	0.5589	1,2,3,4,5	93,820	≥5	7.71	58234.07	0.6208
impMLE (LTA)	0.6911	0.4441	0.4939	1,2,3,4,5,6	69,110	≥6	7.97	55481.51	0.8030
impMLE (LTA)	0.4573	0.4128	0.456	1,2,3,4,5,6,9	45,730	≥7	8.31	45670.55	0.9991
impMLE (LTA)	0.1938	0.3851	0.4516	1,2,3,4,5,6,8,9	19,380	≥8	8.89	24376.16	1.2590
impMLE (LTA)	0.1718	0.3863	0.4468	1,2,3,4,5,6,7,8,9	17,180	≥9	9	21904.5	1.2763
impMLE (PHYS)	0.9382	0.5313	0.5592	1,2,3,4,5	93,820	≥5	8.01	58262.22	0.6211
impMLE (PHYS)	0.687	0.4527	0.4941	1,2,3,4,5,6	68,700	≥6	8.28	55447.77	0.8073
impMLE (PHYS)	0.6015	0.4471	0.4802	1,2,3,4,5,6,10	60,150	≥7	8.40	52853.81	0.8789
impMLE (PHYS)	0.3116	0.4333	0.4712	1,2,3,4,5,6,8,10	31,160	≥8	8.88	36133.14	1.1603
impMLE (PHYS)	0.2755	0.4329	0.462	1,2,3,4,5,6,7,8,10	27,550	≥9	9	32688.08	1.1873
impMLE	1	0.4847	0.5369	1,2	100,000	≥2	6.59	53690	0.5370
impMLE	0.9992	0.484	0.5365	1,2,3	99,920	≥3	6.59	53687.02	0.5374
impMLE	0.9923	0.483	0.5364	1,2,3,5	99,230	≥4	6.60	53991.04	0.5442
impMLE	0.9058	0.4498	0.5134	1,2,3,5,6	90,580	≥5	6.69	55036.41	0.6077
impMLE	0.8567	0.4467	0.5064	1,2,3,5,6,8	85,670	≥6	6.74	55659.8	0.6498
impMLE	0.6353	0.4424	0.4849	1,2,3,4,5,6,8	63,530	≥7	7	53975.09	0.8498

Table 4. Imputation confusion matrix. Most test statistics are calculated using a combination of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) associations. Results from bootstrap imputation of associated NPC class (considering only the most likely class imputed) for 100,000 stands are presented here. Statistics listed here include Matthews correlation coefficient (MC) (Mathews 1975), F-measure (F1) (van Rijsbergen 1979), Rand Index of accuracy (RI) (Rand 1971), and positive and negative predictive value (e.g., PPV and NPV).

	Observed	Not Observed	
Imputed	<i>TP</i>	<i>FP</i>	
	46,312	53,688	PPV=0.4631
Not Imputed	<i>FN</i>	<i>TN</i>	
	48,147	3,450,088	NPV=0.9862
	MC=0.4760	F1=0.4763	RI=0.9722

1. Forest Inventory

Photo: USDA Forest Service

2. Data Layers

- a. Native Plant Community (NPC) plots
- b. Soils
- c. Physiography
- d. More...

3. Imputation Model

- Use rule set and data to identify similar stands.
 - *Rules = Site characteristics (1 + 2b,c,d)*
- Evaluate NPC distribution of matching vegetation plots (2a).
- Assign maximum likelihood classifier.

1. Define Matching Rules

e.g., ID similar stands

2. Tally NPC Observations (2a)

e.g., Rank most likely NPCs

3. Update Inventory Record

4. Ecological Classifications