CURIOUS OR SPURIOUS CORRELATIONS WITHIN A NATIONAL-SCALE FOREST INVENTORY?

Christopher W. Woodall and James A. Westfall

ABSTRACT

Foresters are increasingly required to assess trends not only in traditional forest attributes (e.g., growing-stock volumes), but also across suites of forest health indicators and site/climate variables. Given the tenuous relationship between correlation and causality within extremely large datasets, the goal of this study was to use a nationwide annual forest inventory to determine levels of correlation among a wide array of database fields to aid foresters in separating correlation from causality in comprehensive forest resource assessments. In examining more than 15,000 individual correlations, we found the overwhelming majority (> 85 percent) of correlation coefficients were under 0.1. Site variables (e.g., elevation) had the highest mean correlations, while tree variables (e.g., live aboveground biomass) had the lowest mean correlations with all other variables. Nearly all the high correlations (>0.6) were between variables substantially autocorrelated (e.g., site class code and site index). Given that most correlations within a large-scale forest inventory dataset are very low with the remainder being nonsensical or autocorrelates, finding a highly correlated pair of variables with no apparent autocorrelation deserves further exploration.

INTRODUCTION

For most of the 20th century, forest resource assessments in the United States and abroad were often conducted purposively at small scales using spatially inconsistent sample techniques (i.e., relevé sampling such as stand exams) or conducted periodically at large scales using temporally inconsistent sample techniques (e.g., periodic forest inventory programs in the U.S., Frayer and Furnival 1999). In addition to the lack of spatially and temporally consistent forest inventories, the absence of computing resources available to forest professionals prevented complex forest inventory analyses and resource hypothesis testing. Until the 1990s, the analysis of large-scale forest resource datasets was severely limited to a few analysts with access to inconsistent datasets in computationally limited data management systems.

With the emergence of international agreements focused on the health of forest biomes (USDA 2004) and greenhouse gas accounting, nations have responded by developing nationally consistent forest inventories including numerous variables complementary to traditional tree attributes (e.g., soils and downed dead wood, Perry and others 2009). In addition to field implementation of large-scale forest inventories, data management systems have been developed such that the multitude of data can be rapidly distributed to the public via well-documented web sites. Perhaps never before have forest professionals or the public had access to such large and extensive datasets for exploration of forest resource questions. For example, there are currently 1.1 and 15.0 million records within the plot and tree tables of the U.S. national inventory, respectively (Woudenberg and others 2011). Coupling the millions of inventory records with the hundreds of database fields provides the opportunity to explore numerous facets of forest ecosystems such as fire ecology (Woodall and Nagel 2007), climate change impacts (Woodall and others 2009), forest health (Huebner and others 2009), growth and mortality (Shaw and others 2005), and ownership patterns (Butler and Leatherberry 2005).

With the ability to rapidly assess forest resource attributes using extensive datasets comes the danger of inferring causality from possibly spurious correlations. Given that the U.S. national forest inventory data are publicly available for rapid download, most analyses will be conducted by users not affiliated with the actual data collection or management. Forest professionals have received little guidance on the frequency of high correlations within large-scale forest inventory datasets. Are strong correlations a common occurrence? Does autocorrelation confound many analyses? The goal of this study was to use a nationwide annual forest inventory to determine levels of correlation among a wide array of database fields to help foresters separate correlation from causality in comprehensive forest resource assessments.

METHODS

This study used data exclusively from the national inventory of all U.S. forests. The U.S. Department of Agriculture,

C.W. Woodall, Research Forester, U.S. Department of Agriculture, Forest Service, Northern Research Station, Forest Inventory and Analysis Program, St. Paul, MN 55108
J.A. Westfall, Research Forester, U.S. Department of Agriculture, Forest Service, Northern Research Station, Forest Inventory and Analysis Program, Newtown Square, PA 19073

Forest Service's Forest Inventory and Analysis (FIA) program is charged by Congress with providing an annual inventory of all forest lands. The FIA sampling framework is based on a systematic network of ground plots (Bechtold and Patterson 2005) obtained by dividing the U.S. into a series of 2,400-ha hexagons. Within each hexagon, FIA operates a multi-phase inventory. In phase 1 (P1), land area is stratified using aerial photography or classified satellite imagery to increase the precision of estimates using stratified estimation. In second phase (P2), permanent fixed-area plots are installed in each hexagon when field crews visit plot locations that have accessible forest land. Field crews collect data on more than 300 variables, including land ownership, forest type, tree species, tree size, tree condition, and other site attributes (e.g., slope, aspect, disturbance, land use) (USDA 2009). The plot design for FIA inventory plots consists of four 7.2-m fixed-radius subplots spaced 36.6 m apart in a triangular arrangement, with one subplot in the center. All trees with a diameter at breast height of at least 12.7 cm are inventoried within forested conditions. Within each subplot, a 2.07-m microplot offset 3.66 m from the subplot center is established where live tree seedlings and trees with a d.b.h. between 2.5 and 12.7 cm are inventoried. In addition to the trees measured on these plots, data are also gathered on the condition of the area in which the trees are located (e.g., stand-age class, ownership group, tree-density class). During the third phase of the inventory (P3), forest health indicators are measured on a 1/16th subset of the entire FIA ground plot network. The suite of forest health indicators includes tree crown condition, lichen communities, forest soils, vegetation diversity, down woody material, and ozone injury (Woodall and others In Press).

Using FIA's national database (FIADB version 4.0), we extracted forest inventory data for the most recent inventory in 49 states (currently no inventory available for Hawaii or interior Alaska). Given the multitude of database fields and tables examined in this study, FIA's documented nomenclature will be used in this study (Woudenderg and others 2011). The data extraction was limited to fields in the plot, condition, and tree tables, or variables calculated from those fields (e.g., total tree biomass on a plot): INVYR, STATECD, UNITCD, COUNTYCD, PLOT, PLOT_STATUS_CD, MEASYEAR, MEASMON, MEASDAY, REMPER, KINDCD, DESIGNCD, RDDISTCD, WATERCD, LAT, LON, ELEV, P2PANEL. CONGCD, MANUAL, EMAP_HEX, CYCLE, SUBCYCLE, CONDID, COND_STATUS_CD, RESERVCD, OWNCD, OWNGRPCD, FORTYPCD, FLDTYPCD, MAPDEN, STDAGE, STDSZCD, FLDSZCD, SITECLCD, SICOND, SIBASE, SISP, STDORGCD, CONDPROP_UNADJ, MICRPROP_UNADJ, SUBPPROP_UNADJ, SLOPE, ASPECT, PHYSCLCD, GSSTKCD, ALSTKCD, DSTRBCD1, DSTRBYR1, TRTCD1, TRTYR1, BALIVE, FLDAGE, ALSTK, GSSTK, FORTYPCDCALC, SITETREE_TREE, SITECL_METHOD, CARBON_DOWN_DEAD, CARBON_LITTER, CARBON_SOIL_ORG, CARBON_STANDING_DEAD, CARBON_UNDERSTORY_AG, CARBON_UNDERSTORY_BG, CYCLE2, SUBCYCLE2, TREE, AZIMUTH, DIST, SPCD, SPGRPCD, DIA, DIAHTCD, HT, HTCD, ACTUALHT, TREECLCD, CR, CCLCD, TREEGRCD, CULL, DAMLOC1, DAMTYP1, DAMSEV1, STOCKING, VOLCFNET, VOLCFGRS, VOLBFNET, BOLBFGRS, VOLCFSND, DRYBIO_BOLE, DRYBIO_TOP, DRYBIO_STUMP, DRYBIO_SAPLING, DRYBIO_BG, CARBON_AG, CARBON_BG. All tree-level variables were summed to the plot and condition for live and standing dead trees. These calculated tree-level variables were delineated for live or dead by preceding each variable with a "L" or "D," respectively. Not all fields from the database tables were extracted for this study. Excluded were variables that were not alphanumeric or were a duplication of variables (e.g., secondary and tertiary tree damages). Finally, records were excluded when one or more fields were null. With these constraints, this study's data records totaled 42,617.

Correlations were calculated using SAS's CORR procedure with Pearson's correlation coefficients as the primary output. To assess the distribution of correlations from a large-scale forest inventory, the frequency of correlations from the correlation matrix of all this study's variables was determined. Correlations among the same variables were excluded from the matrix calculation (coefficient=1) for a total of 15,751 correlations. Mean absolute correlations were determined among broad categories of variables according to plot (i.e, plot selection information such as measurement year and county), site (i.e., physiographic information such as elevation and latitude), condition (i.e., stand condition information such as forest type and stand age), and tree (i.e., summed tree attributes such as height and volume). Actual individual correlations were examined when correlation coefficients exceeded 0.7.

RESULTS AND DISCUSSION

In examining of 15,625 individual absolute correlations, we found the overwhelming majority (> 85 percent) to be under 0.1 while less than 1 percent was above 0.5 (Fig. 1). Site variables (e.g., elevation and latitude) had the highest mean correlations (≈ 0.09), while tree variables (e.g., live aboveground biomass) had the lowest mean correlations (≈ 0.05) with all other study variables. Nearly all the high correlations (>0.7) were between variables substantially autocorrelated (e.g., algorithm calculated forest type and field estimated forest type) (Table 1). The remainder of high correlations could be attributed to spurious effects of

database manipulation (e.g., latitude and plot number) or possible curious ecological relationships (e.g., physiographic class and disturbance year).

If variables would be randomly chosen from a strategic-scale forest inventory dataset such as FIA's national inventory, it is extremely unlikely that any appreciable level of correlation would be found. If indeed the correlation exceeded 0.5, then these variables would stand a strong chance of being autocorrelated. Examples of autocorrelation in this study were measurement year and manual number, inventory cycle and kind code, and sum of live tree numbers and sum of distances to live trees. Most of these spurious correlations should be readily identified by even novice inventory analysts. Other spurious correlations, such as longitude and site index base, may take identification by experts in forest inventory databases and sampling designs. Only about a dozen correlations exceeded 0.6 and were ecologically interesting. Physiographic class was strongly correlated with the year of the most recent disturbance, soil organic carbon, understory aboveground biomass, and sum of live-tree board foot gross volume. Poor physiographic sites (i.e., ridge tops) may have shallow soils with little organic soil carbon and may be more prone to disturbances thus reducing their aboveground biomass. It appears as though approaching such large-scale datasets with readily testable ecological hypotheses may be the best method to derive meaningful relationships as opposed to the often spurious results of massive database computations using no *a priori* assumptions.

CONCLUSIONS

Given that most correlations within a large-scale forest inventory dataset are very low with most of the remainder being autocorrelates, finding a highly correlated pair of variables with no apparent autocorrelation is very unlikely. Because all correlations were assumed to be linear in this study, we suggest that non-linear correlations be examined in future studies. With the ever increasing availability of large datasets of ecosystem conditions (i.e., national forest inventories), a tenet can be forwarded: given the extreme rarity of finding highly correlated natural ecosystem variables lacking autocorrelation, when identified their further investigation is warranted.

LITERATURE CITED

Bechtold, W.A.; Patterson, P.L. eds. 2005. Forest Inventory and Analysis national sample design and estimation procedures. Gen. Tech. Rep. SRS-GTR-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 85 p.

Butler, B.J.; Leatherberry, E.C. 2004. America's family forest owners. Journal of Forestry. 102: 4-14.

Frayer, W.E.; Furnival, G.M. 1999. Forest survey sampling designs, a history. Journal of Forestry. 97: 4-10.

Huebner, C.D.; Morin, R.S.; Zurbriggen, A.; White, R.L.; Moore, A.; Twardus, D. 2009. Patterns of exotic plant invasions in Pennsylvania's Allegheny National Forest using intensive Forest Inventory and Analysis plots. Forest Ecology and Management. 257: 258-270.

Perry, C.H.; Woodall, C.W.; Amacher, M.C.; O'Neill, K.P. 2009. An inventory of carbon storage in forest soil and down wood of the United States. In: McPherson, B.J.; Sundquist, E., eds. Carbon sequestration and its role in the global carbon cycle. AGU Special Monograph 183. Washington, DC: American Geophysical Union: 101–116.

Shaw, J.D.; Steed, B.E.; DeBlander, L.T. 2005. Forest Inventory and Analysis (FIA) annual inventory answers the question: What is happening to pinyon-Juniper woodlands? Journal of Forestry. 103: 280-285

U.S. Department of Agriculture Forest Service. 2004. National report on sustainable forests – 2003. FS-766. Washington, DC: U.S. Department of Agriculture, Forest Service. 139 p.

U.S. Department of Agriculture Forest Service. 2009. Forest Inventory and Analysis national core field guide (Phase 2 and 3), version 4.0. Washington, DC: U.S. Department of Agriculture Forest Service, Forest Inventory and Analysis. http://www.fia.fs.fed.us/library/field-guides-methods-proc/ (accessed December, 2009).

Woodall, C.W.; Conkling, B.L.; Amacher, M.C. [and others]. 2010. The Forest Inventory and Analysis phase 3 indicators database 4.0: Description and users manual. Gen. Tech. Rep. NRS-xx. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station.

Woodall, C.W.; Nagel, L.M. 2007. Down woody fuel loadings dynamics of a large-scale blowdown in northern Minnesota. Forest Ecology and Management. 247: 194-199.

Woodall, C.W.; Oswalt, C.M.; Westfall, J.A.; Perry, C.H.; Nelson, M.D.; Finley, A.O. 2009. An indicator of tree migration in forests of the eastern United States. Forest Ecology and Management. 257: 1434-1444.

Woudenberg, S.W.; Conkling, B.L.; O'Connell, B.M. [and others]. 2011. The Forest Inventory and Analysis Database: database description and users manual version 4.0 for Phase 2. Gen. Tech. Rep. RMRS-GTR-245. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Table 1—Matrix of absolute correlation coefficients for all study correlations exceeding 0.7 (live and dead tree variables designated by L and D, respectively)

Variables	Correlations		
	0.7 – 0.8	0.8 – 0.9	0.9 – 1.0
Plot	Lon		
Measyear	Manual		
Lon	Plot		
Manual	Measyear		
Condid		Condprop_ unadj,microprop_unadj, subprop_unadj	
Owncd			Owngrpcd
Owngrpcd			Owncd
Fortypcd	Fldtypcd		
Fldtypcd	Fortypcd		
Siteclcd	Second		
Sicond	Siteclcd		
Conprop_unadj		Condid	Micrprop_unadj, subprop_unadj
Micrprop_unadj		Condid	Condprop_unadj, subprop_unadj
Subprop_unadj		Condid	Conprop_unadj, Micrprop_unadj
Gsstkcd	Carbon_understory_ag		
Dstrbyr1		Carbon_soil_org	
Trtcd1		Carbon_standing_dead	
Trtyr1	L_carbon_bg		
Fldage		L_dist	
Carbon_soil_org		Dstrbyr1	
Carbon_standing_dead		Trtcd1	
Carbon_understory_bg	Gsstkcd		
L_tree	L_azimuth, l_spgrpcd		
L_azimuth	L_tree	L_spcd, l_sprgrpcd	
L_dist		Fldage	
L_spcd		L_azimuth	L_spgrpcd
L_spgrpcd	L_tree	L_azimuth	L_spcd
L_actualht		L_volbfnet	
L_volbfnet		L_actualht	
L_carbon_bg	Trtyr1		
D_tree			D_damtyp1
D_azimuth			D_damsev1
D_dist			D_decaycd
D_damtyp1			D_tree
D_damsev1			D_azimuth
D_decaycd			D_dist

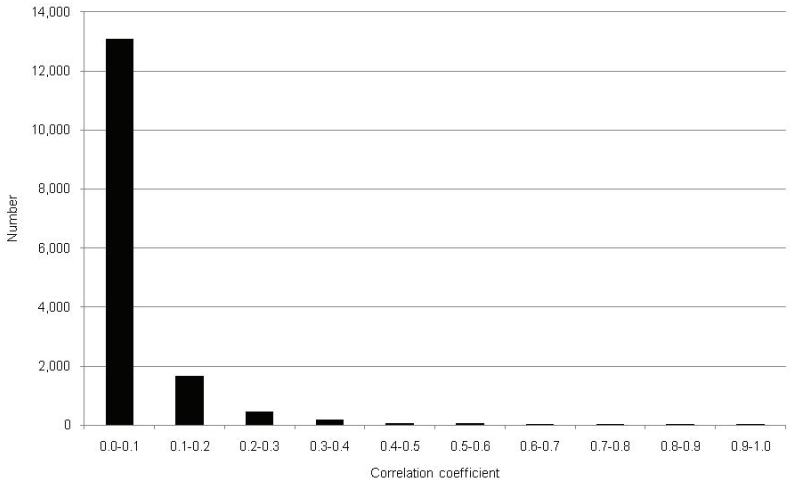

Figure 1—Frequency of absolute correlation coefficients among a multitude of variables sampled during an inventory of U.S. forests.