FIA ESTIMATION IN THE NEW MILLENNIUM[1]

Francis A. Roesch[2]

Abstract—In the new millennium, Forest Inventory and Analysis (FIA) will deliver most of its database information directly to the users over the Internet. This assumption indicates the need for a GIS-based estimation system to support the information delivery system. Presumably, as the data set evolves, it will free FIA and the users from exclusive estimation within political boundaries.

A data set of basal area measurements from a survey unit in Georgia is used to simulate one that might have been obtained had an annual inventory been conducted over a 5-year time interval. The simulated data set was used to investigate various estimators and any potential spatial correlation of basal area. The presence of spatial correlation, coupled with a desire to fulfill user needs to obtain estimates over individually defined elements of the spatial-temporal cube, forms the basis for an argument that a real-time GIS-based estimation system should be developed as the main information delivery vehicle for FIA.

INTRODUCTION

As we approach the next millennium, it is apparent that we cannot consider how we might improve Forest Inventory and Analysis (FIA) estimation without first asking: "What are the major products of FIA likely to be?" That is, through which routes will we deliver the bulk of our inventory information? Most likely, we will deliver most of our information directly from the database over the Internet, not in the paper reports that have historically taken about 2 years to publish. To use the Internet effectively and efficiently, we must build an estimation system to adequately support the delivery of information that is more sensitive to the needs of its users.

Insights into the needs of these users can come from myriad sources, but none so compelling as the reports of the two Blue Ribbon Panels, BRP I (Anonymous 1992) and BRP II (Anonymous 1998). These panels were formed specifically to provide suggestions for improving the FIA program. For instance, a concern over the potential misuse of FIA data resulted in the following statement from BRP I: "To maintain the credibility of the program, FIA, working together with experienced biometricians, must issue clear direction on the scientifically valid uses of FIA data without creating disincentives to innovation and advancement of technology" (Anonymous 1992).

The best way to communicate scientifically valid uses of the information is to develop a system that can provide estimates in as many usable forms as possible. In this manner, FIA will provide scientifically defensible mechanisms from which to make estimates. FIA may still challenge inferences drawn from the estimates, but if the estimates themselves are sound, the scientific community can debate the validity of various resulting inferences.

The second Blue Ribbon Panel reiterated and expanded the recommendations of the first in one specific recommendation:

"Better analysis is necessary for improving customer service. More analysis of FIA data would be useful in improving and increasing customer service. While some FIA customers have the capability and inclination to analyze raw data themselves, other customers rely on outside sources to summarize and analyze the data for them" (Anonymous 1998).

In addition, the second Blue Ribbon Panel charged FIA to "Produce the most current resource data possible."

The overwhelming consensus among panel members was that timeliness of resource data is of paramount importance:

"Strengthening of Forest Service research and expertise in Geographic Information Systems (GIS), and collaboration with other agencies, could deliver immediate benefits. We urge the Forest Service to:

"Reallocate funding within the Forest Service in order to reach the goal of timely resource data established in the first Blue Ribbon Panel report. Fully integrate GIS technology into the inventory process. Aggressively support and promote the annual inventory systems being established in the North Central and Southern FIA units. Based upon results from these efforts, establish a model for annual inventory to be adopted nationwide" (Anonymous 1998).

To fully comprehend the needs of the users, we must first identify those users. FIA users include State foresters, university researchers, National Forest System employees, Forest Service researchers, military bases, other government and State agencies, forest industry, forestry consultants, and members of conservation and environmental groups. Their needs are as diverse as the groups themselves.

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Mathematical Statistician, USDA Forest Service, Southern Research Station, Asheville, NC 28804.

The term “drill-down technology” refers to a database feature that allows a user to view increasing levels of detail as scale increases. It is used extensively in mapping software and GIS packages to deliver the appropriate level of information at varying scales. For example, if one were interested in regional wood supply information, providing individual tree-level data would not be very helpful. On the other hand, region-level information provides little more than background to a State forester interested in a particular county’s available forest resource. A resourceful user can derive region-level information from tree level data from the entire region. However, users are generally supportive of programs that provide information in the most useful forms and at the most appropriate scales. It is apparent that the most efficient aggregator of FIA data into appropriately scaled information will usually be FIA itself. Exceptions do occur with a few special-needs users.

Given the richness of the data that will be available from the annual inventory design, and the power of existing GIS systems, the user of FIA data should not be constrained by boundaries in space or time that have been predefined by FIA. However, given user-defined spatial and temporal constraints, FIA should endeavor to provide estimates in real time. These estimators should be available for as diverse a set of needs as the data will allow. Certainly, the data support investigations into forest amenities and commodities at a wide range of scales, but they can also provide insights into the contributions and effects of forests in wide-ranging areas of interest. Some related areas include studies of pollution, watersheds, and even human behavior.

To increase the usefulness of our information, we must incorporate all of the available improvements in user-interface tools. As a minimum standard, the user should not have to know any variant of Structured Query Language (SQL). This requires that we provide the estimation system in a user-friendly GIS environment.

To ensure the greatest utility of the data, FIA must provide an estimation system robust to an unpredictable and uncontrollable set of click events. This will compel FIA statisticians to reach as deeply into their estimation toolbox as any single previous effort has ever required. A completely different approach will be required if the user is permitted to define areas of interest—say by digitization or by map overlays—and time periods of interest rather than be required to work within strictly defined boundaries in space and time. Given the plethora of information available in the data set, a truly robust system would often have to use estimators that “reach out” to external data and other information sources for support, rather than to rely solely on the FIA data collected within the user-defined, spatial-temporal limits. A GIS-based estimation system has to provide the “best” estimators at any scale of interest within the estimation range. For most attributes, the most intensive scale in the estimation range for FIA data includes areas the size of a large county. However, the relationship of available information to the area delineated varies by the size of the area relative to the sample, the variable of interest, and the period of interest. Large areas require only the usual sample estimates when sample sizes within the area and period are adequate, while small areas require the use of supplementary information from outside of the area or period.

METHODS

Before an estimation system can be incorporated into a production system, its individual components, as well as the relationships between those components, must be thoroughly tested. This leaves us with the problem of testing a large, potentially complex, estimation system prior to the availability of the data. The approach we used was to manufacture a reasonably believable 5-year series of data by projecting data from a single year backward and forward 1 and 2 years. The data from FIA’s Survey Unit 1 in Georgia, collected in 1989 and 1996, were used to establish individual-tree basal area projection equations, mortality and harvest probabilities, and proportions by forest type, dominant species, and age class. These functions were then applied to the 1996 tree-level data to project it backward 1 and 2 years and forward 1 and 2 years, simulating tree data for 5 consecutive years on 2,353 plots. The survey unit consists of 35 counties, which were grouped into 5 contiguous 7-county groups for part of this study. This data set was considered to represent the “truth” for each of the years 1994 through 1998. Figure 1 graphs the “true” mean basal areas per acre. We define the “current truth” as the state of this simulated population in 1998.

The sample plots for the FIA Annual Inventory sample design are located in a systematic triangular grid consisting of five interpenetrating panels. One panel is measured each year for five consecutive years, after which the panel measurement sequence reinitiates. If panel 1 was measured in 1998, it will also be measured in 2003, 2008, and so on. Panel 2 would be measured in 1999, 2004, 2009, and every five years thereafter.

To mimic the systematic FIA Annual Inventory design, spatial coordinates of the plots were used to assign plots to panels, a panel being a single year’s measurement. Therefore, the simulated FIA Annual Inventory sample consisted of approximately one-fifth of the plots for each year.

A preliminary study investigated specific applications of two general methods for combining the multiyear data from the FIA annual inventory design to form current estimates for small areas. The two general methods are (1) the simple moving average estimator (MAE), and (2) a globally defined mixed estimator (ME) applied locally. Two variations of the mixed estimator method (ME1 and ME2) are compared to each other as well as to the assumed default estimator (MAE). Assume that one and only one full series of observations is available so that all five panels have been measured once. “Current” is defined as the measurement time of the last panel (panel 5).

MAE pools the latest five panels measured, under the assumption that no time trend exists at the observed scale. As some variables of interest will violate this assumption

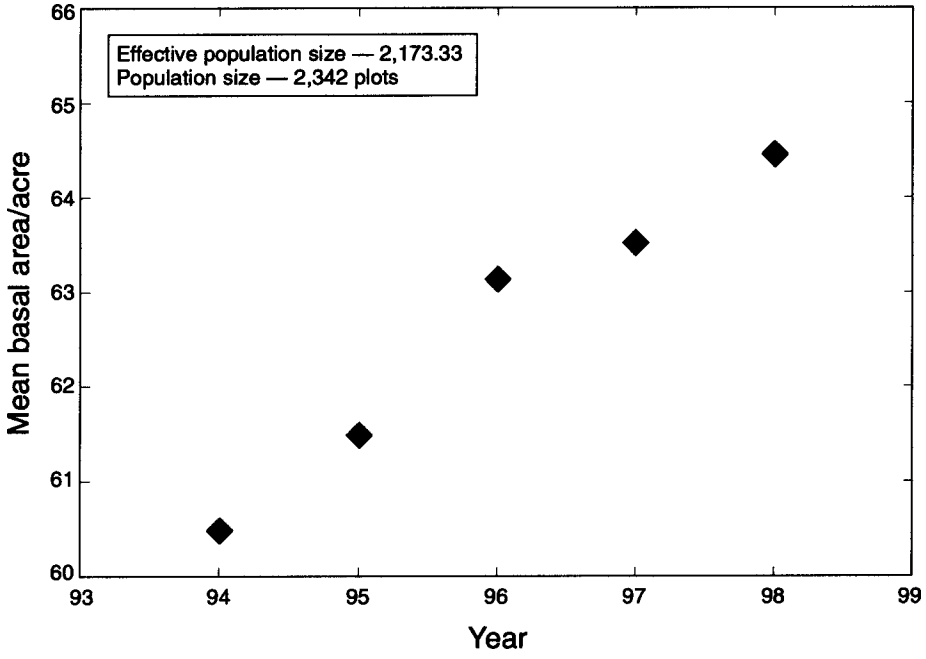

Figure 1—Survey unit "true" mean basal area per acre by year.

over the measurement interval, it is important to determine valid applications of this simple estimator.

Modeling an existing temporal trend becomes important when the objective is to estimate the time-specific value of some forest attribute, e.g. current volume or basal area per acre. When a temporal trend does in fact exist, MAE will have the tendency to mask the very trends that the FIA annual inventory design was intended to evaluate. Therefore, we explored the mixed estimator because it can recognize and efficiently utilize the time-series nature of the five-panel sample.

If we seek the estimate for a variable at a specific time, let:

X_{ijt} = the per-acre value observed at plot i in county j

($i = 1,..., n_j$, $j = 1,...,J$), and time t ($t = 1,...,5$),

A_{ijt} = the area in acres sampled at plot i in county j

($i = 1,..., n_j$, $j = 1,...,J$), and time t ($t = 1,...,5$), and

A_P = the fixed plot area.

When no time trend is present, the sample area weighted mean for the five-panel series provides the best estimator of a per-acre value (V):

$$\hat{V}_{MAE} = \frac{1}{A}\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{A_{ijt}}{A_P}X_{ijt} \qquad (1)$$

where:

$$A = \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{i=1}^{n_j} A_{ijt}$$

We tested two variations of mixed estimation for current basal area. Each variation applies global (survey unit) results of the mixed estimation methodology to subareas within the survey unit, under the assumption that the sample will often be too small for a direct application of mixed estimation to the subareas. In both variations, we used mixed estimation at the survey unit level to choose from the three simple models discussed by Van Deusen (1999), and to find the maximum likelihood estimate of the weighting parameter p. The models were (1) a straight line with a slope of zero, (2) a straight line of any slope, and (3) a quadratic. In the first variation (ME1), we fit the chosen model and level of p at the lower levels (i.e. county and county group). In the second variation (ME2), we fit the chosen model at the survey unit level to predict an overall $\hat{\beta}$ (a Tx1 vector described below, where T is the number of years in the sample, usually equal to 5). This leads directly to a simple updating vector $\mathbf{U}$, found by multiplying the inverse of each element of $\hat{\beta}$ by the fifth element of $\hat{\beta}$. Then:

$$\hat{V}_{ME2} = \left(1'\mathbf{A}_T\right)^{-1}\left(\left(DIAGRV\left(\mathbf{A}_T\right)\right)\mathbf{V}_T\right)'\mathbf{U}$$

where:

$\mathbf{A}_T$ = a Tx1 vector of total area sampled at each time,

$\mathbf{V}_T$ = a Tx1 vector of basal area estimates for each time,

$\mathbf{1}$ = a Tx1 vector of ones, and

$\mathrm{DIAGRV}\left(\mathbf{A}_T\right)$ = a function that places a Tx1 vector $\mathbf{A}_T$ into the diagonal of a TxT matrix of zeroes.

We then evaluated the estimators for how well they predicted the "true" county level and county group level basal areas for 1998 from the 1994 to 1998 sample, under a squared error loss function. Initially, we conducted a case study yielding a unique solution for the moving average estimator (MAE) and each variation of mixed estimation (ME1 and ME2). The squared error calculated for these methods is simply the mean of the squared difference of each estimate by county and county group from the truth for that county or county group.

Finally, we performed a simulation, assuming that spatial correlation between plots was unimportant. The plots were randomly rearranged 1,000 times and then grouped into simulated, approximately equally sized, "counties." We varied the number of counties from 10 to 50 to see what effects sample size would have on the ranking of the estimation approaches. When the plots were grouped into 10 counties, there was an average of 235.3 plots in each county, (actually 235 plots in 7 counties and 236 plots in 3 counties). At the other extreme, when there were 50 counties, there were 47 plots in 47 counties and 48 plots in 3 counties. We calculated the mean difference and mean squared difference from the "truth" over the 1,000 random arrangements of the 2,353 plots. We defined the "truth" as the population mean of each simulated county at time 5.

The simulation results led to the suspicion that the assumption of spatial independence between plots was weak. Therefore, in an attempt to detect spatial trends, we performed median polishes of the "true" population plot data for 1998 aggregated at five different scales of a square grid (50, 40, 30, 20, and 10 miles on a side). We conducted the median polishes in the cardinal directions (north-south and east-west). At two scales, a strong north-south trend was indicated. The cell sizes for the first of these scales were slightly larger than the average county size (a square grid with 30 miles on a side), resulting in 31 filled cells. The second scale was 20 miles on a side, resulting in 57 filled cells. The results for 50, 40, and 10-mile grids are not presented because they did not show any spatial trends. Subsequent to the median polishes, we calculated the variograms for the 30 and 20-mile grids of both the original data and the residuals.

RESULTS

For the case study, table 1 shows the mean difference from the truth over all counties and county groups for MAE, ME1, ME2, and the mean of panel 5 (P5M). Table 2 shows the corresponding mean squared differences. The panel 5 mean is included because panel 5 is the portion of the sample that observes only the population partition of interest (that is, tree basal areas during 1998). In the case study, the mean difference is not a true measure of model bias, but can be an indication of model bias. Note that two of the estimators have roughly the same mean difference at both the county and county group levels, leading us to suspect that the respective levels may reflect the true level of bias in these estimators. Of these two, MAE shows the largest absolute difference. Due to the increasing trend in the variable of interest, all values for the moving average were low. The magnitude of the absolute mean difference

Table 1—Mean difference—case study

Estimator	County	County group
Moving average estimator	-2.026	-1.919
Panel 5 mean	2.293	.024
Mixed estimator, variation 1	-2.367	.027
Mixed estimator, variation 2	.078	.159

Table 2—Mean squared difference—case study

Estimator	County	County group
Moving average estimator	12.586	4.305
Panel 5 mean	314.065	8.963
Mixed estimator, variation 1	98.350	2.470
Mixed estimator, variation 2	9.128	.513

is close to zero for ME2. When going from the county to the county group level, the large reduction in magnitude of absolute mean difference for the other two estimators appears to be more a result of decreasing variance than of bias. Of course, because P5M is design unbiased and does not rely on a time dependent model, we know that this is the case for P5M.

In table 2, ME2 shows the lowest mean squared differences overall. In addition, ME1 has a higher variance than MAE and ME2 at the county level, because the sample sizes were too small at the county level to fit the model. Two observations support this statement. First, ME1 behaves better at the county group level than at the county level. Second, ME2, in which the model was fit at the survey unit level and then applied at the lower levels, works well even at the county level.

The second part of the study, the simulation in which we randomly rearranged the plots, has led to unexpected, albeit explainable results. The top graph in figure 2 shows the mean squared difference from the truth for the 1,000 random arrangements of the 2,353 plots after being grouped into 10 to 50 counties; the bottom graph gives the corresponding mean differences. Note that although the MAE of time 5 basal area still displays the expected bias, it now compares favorably, in terms of mean squared error, with ME2. ME2 can be expected to work best if the individual county basal areas at times 1 through 4 have the same values relative to the county basal areas at time 5 as occurs globally over the entire survey unit. In a heterogeneous population, this condition is more likely to occur if similar plots are spatially collocated. ME1 requires that the

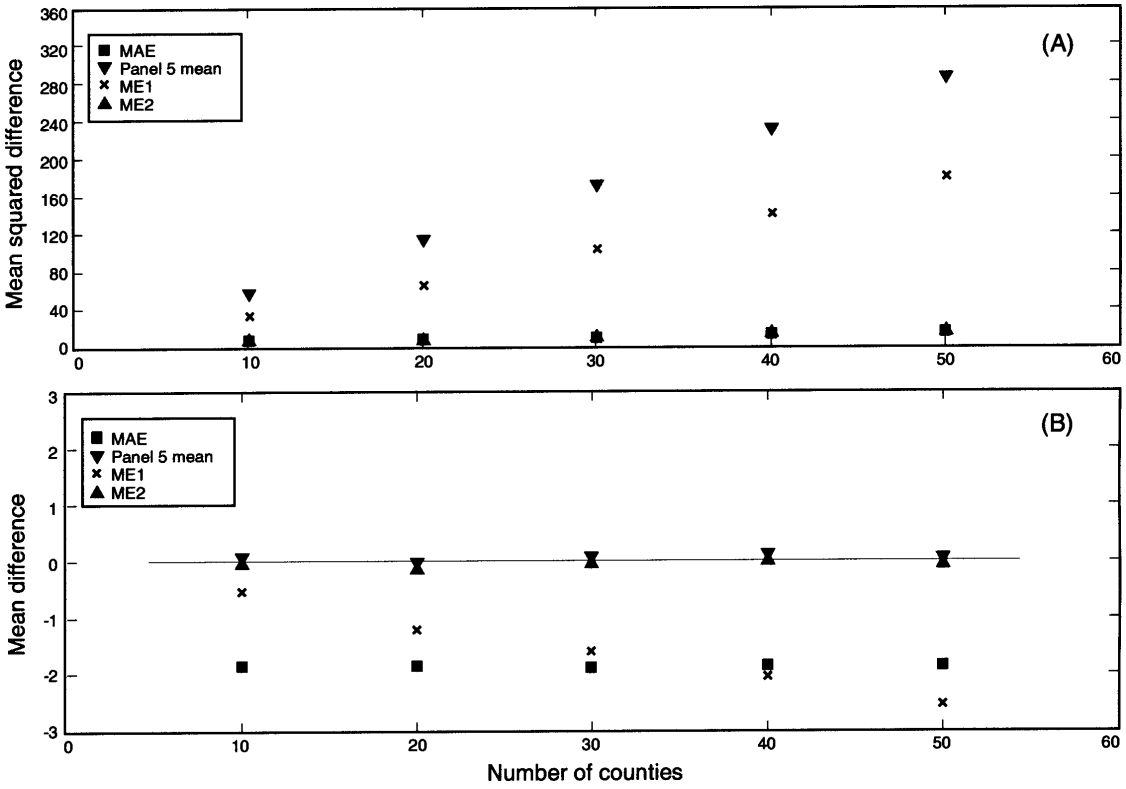

Figure 2—Mean squared difference (A) and mean difference (B) from the truth for 1,000 random arrangements of the 2,353 plots after being grouped into 10 to 50 counties; i.e., sample size per county is decreasing from left to right.

selected model be fit at the county level. This would be advantageous if plots within counties were more homogeneous than between counties, and if there were a sufficient number of plots in each county. These observations have led to a search for spatial trends in the data.

In the top graph, figure 3 gives the aggregated mean population basal areas for the data within grid cells of 20 miles on a side; in the bottom graph, it gives the residuals, row, column, and all effects following a median polish of this data. Tukey (1977) and Cressie (1991) explain the median polish (also known as median sweep). Figure 4 gives the corresponding information following a coarse mapping with 30-mile grid cells. In the bottom graph of both figures, the row effect (far right column, save for the "all" effect at the bottom) is a large, positive number at the top and a not-quite-as-large, negative number at the bottom. Although neither vector strictly decreases from top to bottom, a trend does appear likely.

The top graph of figure 5 shows the classical estimates, as well as the Cressie-Hawkins robust estimates (Cressie and Hawkins 1980), of the north-south variograms for the data in the top graph of figure 3. The bottom graph of figure 5 displays the corresponding estimates for the residuals in figure 3. Likewise, figure 6 provides the same estimates for the data in figure 4. Figure 5 illustrates the classic argument that the median polish removes spatial correlation from the data, as the estimated variograms of the residuals are decidedly flatter than those of the data. At first blush, figure 6 seems to give quite the opposite impression; that is, unless one ignores the values for the lag of six (equal to 180 miles). It is appropriate to ignore this lag since only a single observation supported it and one end of the interval happens to be in a row with only two observations. Ignoring the lag 6 values, we see that the plots for the residuals are slightly flatter than the plots for the data. In toto, figures 3 through 6 show that there is a north-south trend observable at scales greater than or equal to 100 miles.

CONCLUSIONS

The FIA annual inventory design will provide a set of sample observations of forest attributes that is thoroughly diffused through space and time. This will allow estimation of forest attributes for an almost-infinite set of subdomains of interest. FIA cannot provide this extremely large set of potential estimates; however, FIA could and should provide a reasonable set of tools within an estimation system to users accessing the data over the Internet. Such a system would be more useful if it made reasoned use of data from outside the domain of interest (i.e. the space-time cube defined by a user) when that domain of interest is too small to contain enough observations for the usual sample-based estimators.

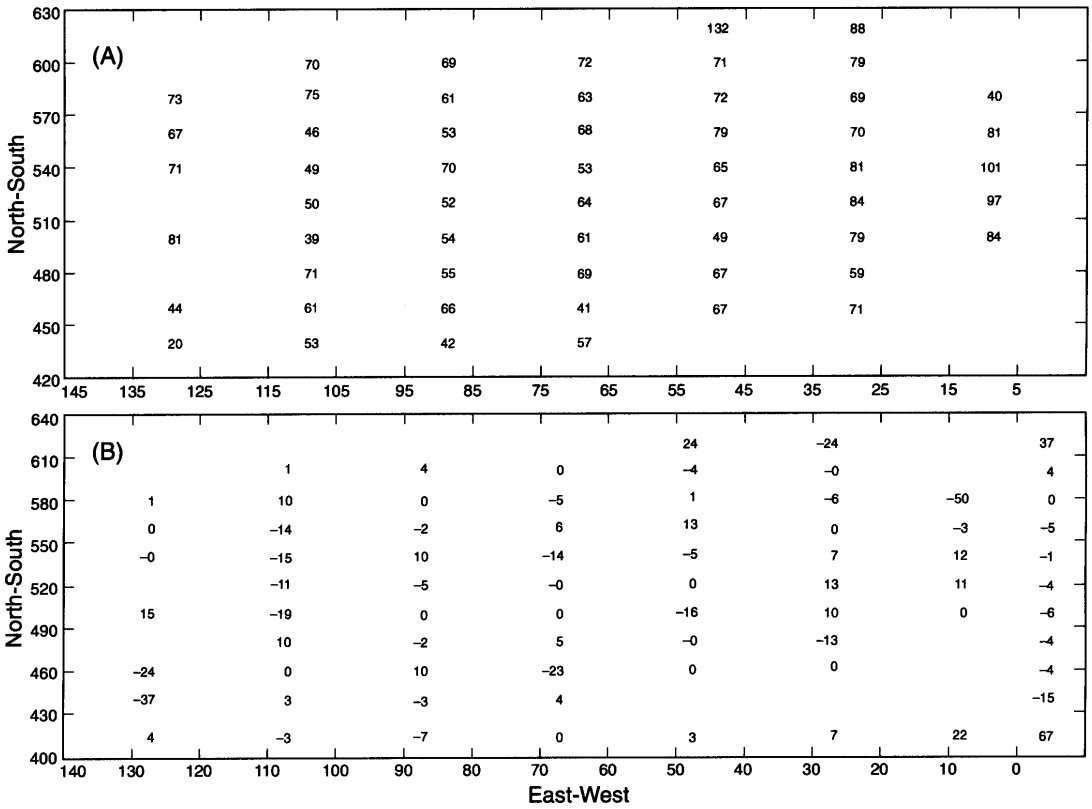

Figure 3—Aggregated mean basal area (A) of the "true" population simulated from Georgia, Survey Unit 1 Forest Inventory and Analysis data following a coarse mapping with a grid size of 20 miles on a side, plotted by an arbitrary coordinate system. The bottom graph (B) shows the overall effect (bottom right), the column effects (remainder of the bottom row), the row effects (remainder of the right column), and the residuals (remaining values) following a median polish of the data in the top graph.

This study examines methods of making estimates over a smaller domain than the sample within that domain will actually support. The methods that use outside information in different ways, MAE, ME1, and ME2, yield substantial improvement in terms of squared error loss over P5M. None of the alternative estimators, as applied to the small-areas, however, can be shown to be design unbiased. In the presence of increasing or decreasing trend, the alternatives to the simple moving average have the potential of being model unbiased. For basal area, and presumably all variables that are likely to exhibit trends over the 5-year measurement period, even simplistic approaches to modeling the trends can result in significant reductions in MSE over the simple moving average.

These estimators (MAE, ME1, and ME2) use the same information in different ways. That information comes only from the FIA annual inventory data, although 80 percent comes from outside of the domain of interest. Therefore, comparisons between the methods are direct. On the other hand, some methods that we have not discussed here benefit from a rich history of external growth and yield research. Mixed estimation, in general, represents a much lower investment in human resources both initially and in the long term than common industrial methods, which use growth and mortality equations to update plot data. This latter approach would be difficult for FIA to use because appropriate growth models do not exist for many condition classes of interest, and those that do exist would have to undergo thorough testing for use in this context. In addition, to ensure that the forest populations are not moving away from those upon which the models were built, the growth model predictions would have to be constantly monitored.

There are at least two ways to view any differences between Part 1, the case study, and Part 2, in which the plots were randomly rearranged 1,000 times. Conducted over a broader range of conditions, the simulation, on the one hand, should be considered a more robust test of the behaviors of the respective estimators. On the other hand, the simulation disfavors estimators that draw strength from spatial correlation, if that correlation exists in real populations. Any spatial correlation inherent in the data remained intact in the case study but not in the simulation. The results support this second viewpoint on a number of fronts. For instance, the moving average estimator moved up in ranking during the simulation relative to the case study. Since spatial correlation would lead to stronger time trends within counties, and the moving average estimator would be at a disadvantage in the presence of a time trend, a simulation ignoring potential spatial correlation might

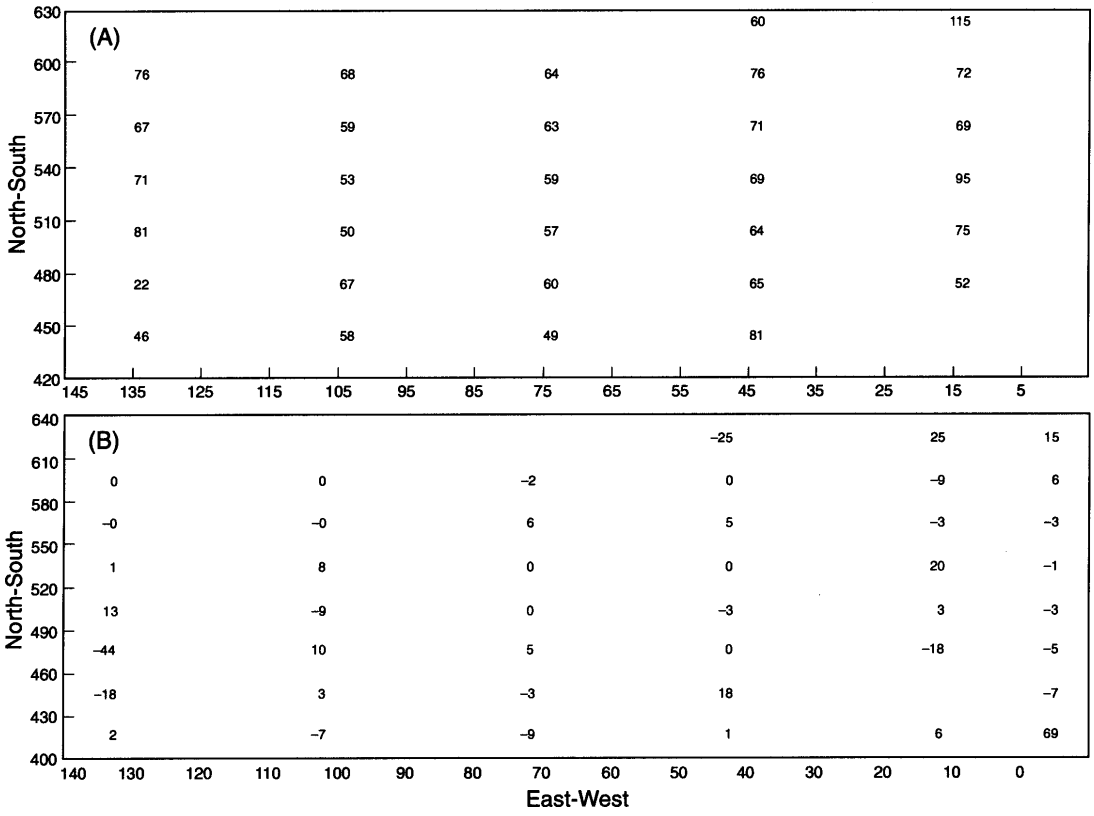

Figure 4—Aggregated mean basal area (A) of the "true" population simulated from Georgia, Survey Unit 1 Forest Inventory and Analysis data following a coarse mapping with a grid size of 30 miles on a side, plotted by an arbitrary coordinate system. The bottom graph (B) shows the overall effect (bottom right), the column effects (remainder of the bottom row), the row effects (remainder of the right column), and the residuals (remaining values) following a median polish of the data in the top graph.

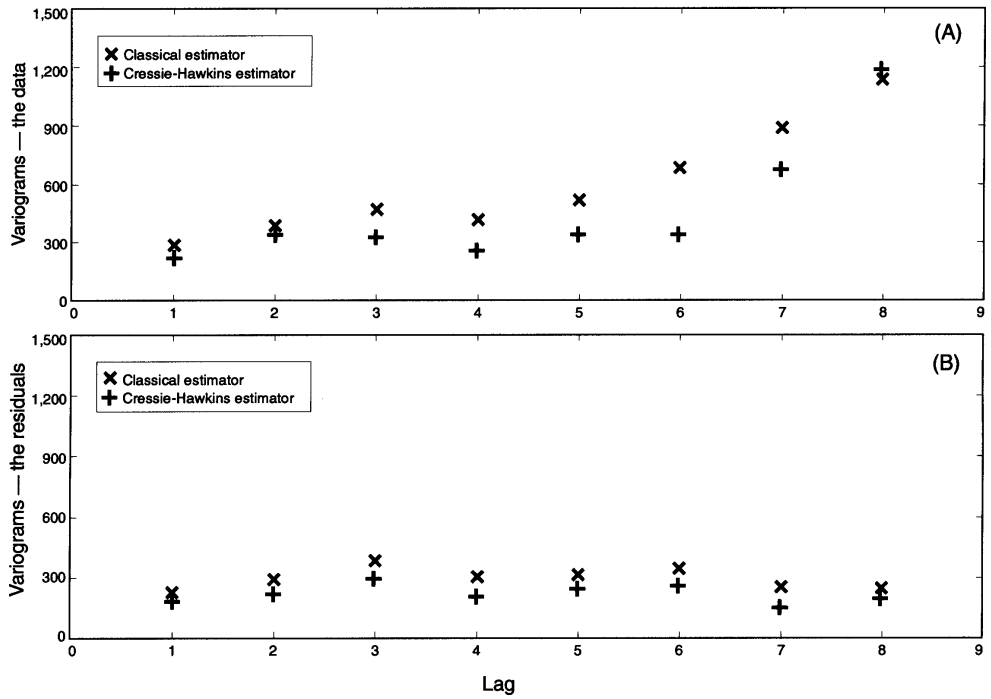

Figure 5—Classical and Cressie-Hawkins robust estimators of the north-south variograms for a grid size of 20 miles on a side, for the aggregated data (A) and the residuals (B) following a north-south, east-west median polish.

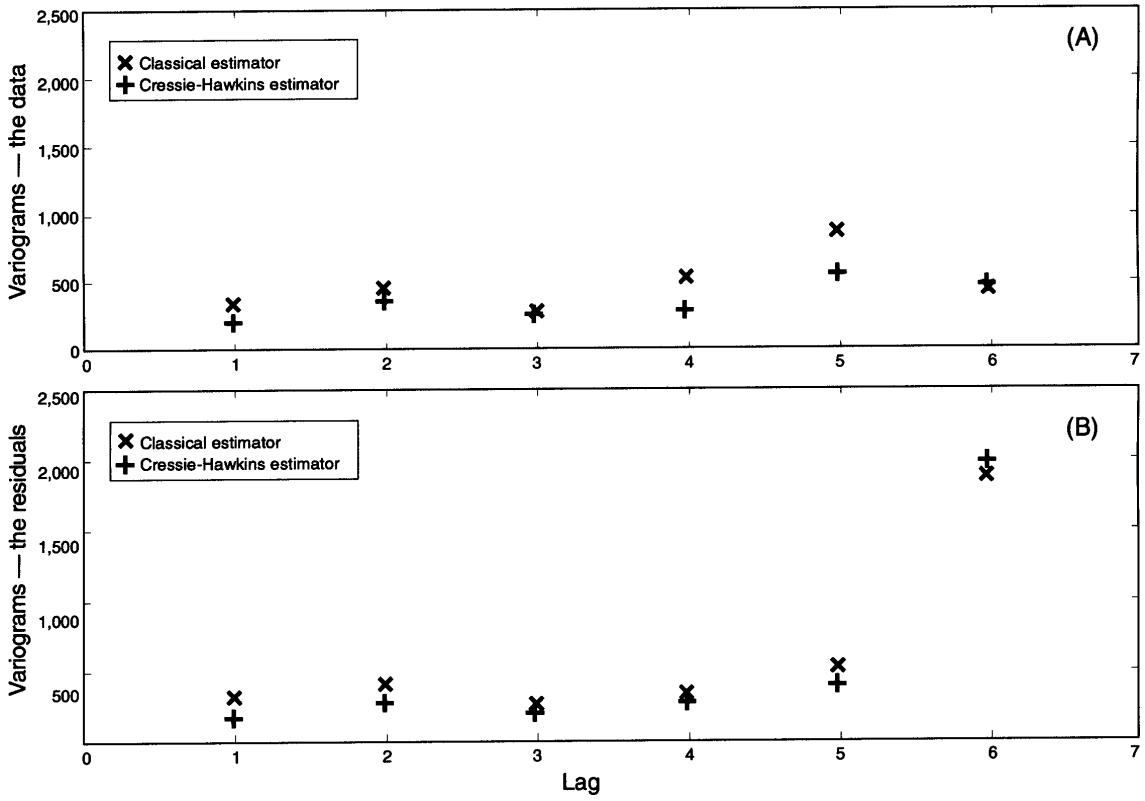

Figure 6—Classical and Cressie-Hawkins robust estimators of the north-south variograms for a grid size of 30 miles on a side, for the aggregated data (A) and the residuals (B) following a North-South, East-West median polish.

garble any time trend enough to favor the moving average estimator. Similarly, if a strong global time trend existed, in the presence of strong spatial correlation at the county level, the two applications of mixed estimation would benefit. Therefore, we should not be surprised if they fare better in the case study than in this particular simulation.

The spatial analysis established that the basal area data did contain spatial correlation at relevant scales. Other survey units, of a similar size and diversity, could also exhibit spatial trends for this and probably other variables. Therefore, modeling for both the potential spatial trends as well as the potential temporal trends within survey units could benefit small-area estimates. This gives further credence to the call to FIA for the development of a GIS based estimation system with the ability to adapt to user-defined areas and periods of interest.

REFERENCES

Anon. 1992. Report of the blue ribbon panel on forest inventory and analysis, Washington, DC. Available from: Frank Roesch, USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, NC.

Anon. 1998. Report of the second blue ribbon panel on forest inventory and analysis, Washington, DC. Available from: Frank Roesch, USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, NC.

Cressie, N. 1991. Statistics for spatial data. New York: John Wiley. 900 p.

Cressie, N.; Hawkins, D. 1980. Robust estimation of the variogram, I. Journal of the International Association for Mathematical Geology. 12: 115–125.

Tukey, J. 1977. Exploratory data analysis. Reading, MA: Addison-Wesley. 688 p.

Van Deusen, P. 1999. Modeling trends with annual survey data. Canadian Journal of Forest Research. 29(12): 1824–1828.