A DATABASE STRATEGY FOR NEW VARIABLES

B. Tyler Wilson, Ali Conner, Glenn Christensen, John Shaw, Jason Meade, and Larry Royer

ABSTRACT

The introduction of new variables into the annual inventory system of the U.S. Forest Service's Forest Inventory and Analysis (FIA) program can create issues with population estimates since evaluations (or expansion factors) based on a full cycle's worth of data should not be used with new data that have not been collected for a full cycle. This manuscript provides guidance on how to manage evaluations within the National Information Management System Compilation System (NIMS-CS) when new variables/attributes are added to the Forest Inventory and Analysis annual inventory.

INTRODUCTION

In an attempt to be responsive to the changing needs of its users, the U.S. Forest Service's Forest Inventory and Analysis (FIA) program sometimes begins collecting new variables on field plots in mid-cycle. For example, to bring our definition of forest into alignment with international standards, FIA has begun collecting tree canopy cover data on all phase 2 field plots. Because only a fraction of field plots are sampled annually (i.e., 10-20 percent, depending upon the state), there will be a period of time that elapses before a full cycle of data is available for these new variables (i.e., 5-10 years). For the purposes of population estimation (Bechtold et al. 2005), all plots collected within a stratum of an estimation unit for an evaluation period are assigned the same plot expansion factor. Expansion factors are computed by dividing the acreage of each stratum within an estimation unit by the numbers of plots sampled in the stratum for the evaluation period. For new variables that have not yet been collected for an entire cycle, the expansion factors associated with a full cycle of plots will be incorrect. Therefore, FIA needs a strategy to handle these differences in plot expansion factors amongst variables prior to the completion of the first full cycle of sampling.

ORGANIZATIONAL STRUCTURE

The first step in the solution was to identify FIA staff to approach the problem from the perspective of each of the regional programs and the functional areas impacted: information management, techniques research, and analysis. Information management staff process the data collected in the field and maintain the FIA database. Staff of techniques research ensure that the field sample is collected and compiled in such a way to permit meaningful population estimation. Analysts interpret the population estimates and produce annual and 5-year reports on the status of forests in their region of expertise. The authors of this manuscript represent each of these regions and functional areas and were identified as the task team. The second step was for the team to develop and analyze a small set of alternative solutions to the problem. In the final step, the team recommended a strategy for handling new variables to the program managers. In this paper, we present a synopsis of the team's analysis and recommendation.

ALTERNATIVE STRATEGIES

The team proposed and examined three alternative strategies for handling new variables. The first approach is to wait until a full cycle of plots is collected before reporting on a new variable. The benefit of this approach is that it requires no additional effort beyond that required to collect, compile, store, analyze, and report these new variables. However, there are a few drawbacks. Waiting until a complete cycle has been collected would result in unacceptable delays (i.e., between 5 years and 10 years) prior to reporting. This would foster the perception of "gate-keeping" by FIA and would not enhance recent efforts by the organization to promote transparency in its methods.

B. Tyler Wilson, Research Forester, Northern Research Station, Saint Paul, MN 55108
Ali Conner, Supervisory, IT Specialist, Southern Research Station, Knoxville, TN 37919
Glenn Christensen, Forester, Pacific Northwest Research Station, Portland, OR 97205
John Shaw, Forester, Rocky Mountain Research Station, Ogden, UT 84401
Jason Meade, Forester, Southern Research Station, Knoxville, TN 37919
Larry Royer, IT Specialist, University of Nevada, Las Vegas, NV 89154

Table 1—Summary of worked example

RPT YR	Data Example	EVAL_GRP	EVALID	EVAL_TYP	Avg. Exp. Factor	# of Plots
2005	AL8_12345 (2001, 2002, 2003, 2004, 2005) (full cycle data)	012005	010550	EXPALL	6,000	4,000
			010551	EXPCURR	6,000	4,000
				EXPVOL	6,000	4,000
			010553	EXPGROW	6,000	4,000
				EXPMORT	6,000	4,000
				EXPREMV	6,000	4,000
	New Variable (A) AL8_5 (2005)	012105	010510	EXPALL	30,000	800
			010511	EXPCURR	30,000	800
				EXPVOL	30,000	800
2006	AL8_2345 + AL9_1 (2002, 2003, 2004, 2005, 2006)	012006	010650	EXPALL	6,000	4,000
			010651	EXPCURR	6,000	4,000
				EXPVOL	6,000	4,000
			010653	EXPGROW	6,000	4,000
				EXPMORT	6,000	4,000
				EXPREMV	6,000	4,000
	Variable (A) AL8_5 + AL9_1 (2005, 2006)	012206	010620	EXPALL	15,000	1,600
			010621	EXPCURR	15,000	1,600
				EXPVOL	15,000	1,600
2007	AL8_345 + AL9_12 (2003, 2004, 2005, 2006, 2007)	012007	010750	EXPALL	6,000	4,000
			010751	EXPCURR	6,000	4,000
				EXPVOL	6,000	4,000
			010753	EXPGROW	6,000	4,000
				EXPMORT	6,000	4,000
				EXPREMV	6,000	4,000
	Variable (A) AL8_5 + AL9_12 (2005, 2006, 2007)	012307	010730	EXPALL	10,000	2,400
			010731	EXPCURR	10,000	2,400
				EXPVOL	10,000	2,400
2008	AL8_45 + AL9_123 (2004, 2005, 2006, 2007, 2008)	012008	010850	EXPALL	6,000	4,000
			10851	EXPCURR	6,000	4,000
				EXPVOL	6,000	4,000
			010853	EXPGROW	6,000	4,000
				EXPMORT	6,000	4,000
				EXPREMV	6,000	4,000
	Variable (A) AL8_5 + AL9_123 (2005, 2006, 2007, 2008)	012408	010840	EXPALL	7,500	3,200
			010841	EXPCURR	7,500	3,200
				EXPVOL	7,500	3,200
	New Variable (B) AL9_3 (2008)	012108	010810	EXPALL	30,000	800
			010811	EXPCURR	30,000	800
				EXPVOL	30,000	800
2009	AL8_5+AL9_1234 5 panels **inc Var (A)** (2005, 2006, 2007, 2008, 2009)	012009	010950	EXPALL	6,000	4,000
			010951	EXPCURR	6,000	4,000
				EXPVOL	6,000	4,000
			010953 **No Var A**	EXPGROW	6,000	4,000
				EXPMORT	6,000	4,000
				EXPREMV	6,000	4,000
	Variable (B) AL9_34 (2008, 2009)	012209	010920	EXPALL	15,000	1,600
			010921	EXPCURR	15,000	1,600
				EXPVOL	15,000	1,600
2010	AL9_12345 (2006, 2007, 2008, 2009, 2010)	012010	011050	EXPALL	6,000	4,000
			011051	EXPCURR	6,000	4,000
				EXPVOL	6,000	4,000
			011053	EXPGROW	6,000	4,000
				EXPMORT	6,000	4,000
				EXPREMV	6,000	4,000
	Variable (A) AL9_5 (2010)	012110	011013	EXPGROW	30,000	800
				EXPMORT	30,000	800
				EXPREMV	30,000	800
	Variable (B) AL9_345 (2008, 2009, 2010)	012310	011030	EXPALL	10,000	2,400
			011031	EXPCURR	10,000	2,400
				EXPVOL	10,000	2,400

The second approach is to modify the code in the reporting tools that calculates estimates to adjust stored expansion factors dynamically to account for incomplete cycles of data for new variables. The benefit is that the underlying data and database structure would remain unchanged. But this approach also has several issues. It would require some complex programming to encode the necessary logic. Furthermore, such business logic is most appropriately stored in the National Information Management System Compilation System (NIMS-CS) , along with all of the other compilation procedures. Finally, the sampling errors of estimates produced using less than a full cycle of plots would be larger because of the smaller numbers of plots in the sample, though this problem goes away once a full cycle of plots is collected.

The third approach is to create separate evaluations for new variables within the NIMS-CS. The benefits are that the database structure would not require any modification other than the additional records to be created, which is true of any evaluation. The drawbacks are that this approach requires slightly more compilation time and the need to maintain more records in the population (POP) tables, and thus possible confusion for which evaluation to chose. It also results in larger sampling errors before the collection of a full cycle of plots. However, both of these problems are eliminated once the cycle for the new variable is complete. Because of the simplicity of the approach and the benefit of reporting on new variables prior to the collection of a full cycle of data, in spite of slightly larger sampling errors in the interim, the third approach was recommended for handling new variables. The application of this approach is illustrated in the next section using a specific example.

EXAMPLE USING RECOMMENDED STRATEGY

For the purposes of clarity, let us make a few simplifying assumptions for the example. Assume that FIA will start collecting new variables on phase 2 plots in Alabama, which is on a 5-year cycle, with 800 plots sampled in an inventory year and 4,000 plots sampled in a full cycle. Assume that the average expansion factor per plot is 6,000 at base sampling intensity for a full cycle of 5 inventory years, giving a total of 24 million acres sampled.

We will assume that a full cycle of data is available for Alabama in 2005, which is comprised of 5 panels of plots sampled during inventory years 2001, 2002, 2003, 2004, and 2005. We will designate an evaluation group (12005) as a reference to this set of plots. This evaluation group includes three evaluations: EXPALL (10550), EXPCURR (10551), and EXPGRM (10553). The EXPALL evaluation includes all of the plots, whether sampled or not. Plots may not be sampled because of hazardous conditions or denied access. The EXPCURR evaluation includes only those plots that were sampled, either via a field or office visit. The EXPGRM evaluation includes only those plots that were sampled at two points in time, thus allowing the calculation of the components of change, broadly categorized as growth, removals, and mortality.

In this example, new variable A is introduced for the 2005 field season. In order to account for the fact that this variable has been collected on only 800 plots and therefore requires a different expansion factor (24 million acres / 800 plots = 30,000 acres/plot) than those variables collected on a full cycle of plots, a second evaluation group is needed (12105). This evaluation group will include two evaluations: EXPALL (10510) and EXPCURR (10511), corresponding to the first two evaluations in the first evaluation group, but using only one panel of plots.

In 2006, all variables are collected on the plots in the panel, including new variable A. For 2006, Alabama again requires two evaluation groups (12006, 12206). Evaluation group 12006 includes plots from 5 inventory years (2002-2006). This evaluation group is comprised of three evaluations: EXPALL (10650), EXPCURR (10651), and EXPGRM (10653). Evaluation group 12206 is created for new variable A, which has now been collected in inventory years 2005 and 2006 on 1,600 plots. This works out to an expansion factor of 15,000 acres/plot. This evaluation group is comprised of two evaluations: EXPALL (10620) and EXPCURR (10621), corresponding to the first two evaluations in the first evaluation group, but using only two panels of plots.

New variable B is introduced in 2008. Because there are different sampling intensities for the original variables, new variable A, and new variable B, Alabama 2008 requires three evaluation groups (12008, 12408, 12108). Evaluation group 12008 includes plots from 5 inventory years (2004-2008). It consists of three evaluations: EXPALL (10850), EXPCURR (10851), and EXPGRM (10853). Evaluation group 12408 is created for new variable A, which has now been collected on approximately 3,200 plots in the inventory years 2005-2008. This works out to an expansion factor of 7,500 acres/plot. This evaluation group is comprised of two evaluations: EXPALL (10840) and EXPCURR (10841), corresponding to the first two evaluations in the first evaluation group, but using only four panels of plots. Evaluation group 12108 is created for new variable B, which has been collected on approximately 800 plots in inventory year 2008, for an average expansion factor of 30,000 acres/plot. It consists of two evaluations: EXPALL (10810) and EXPCURR (10811), corresponding to the first two evaluations in the first evaluation group, but using only one panel of plots.

In 2009, Alabama requires only two evaluation groups (12009, 12209). The reason for this is that new variable A has now been collected for a complete cycle of plots (2005-2009). Evaluation group 12009 includes all of the original variables, as well as new variable A. This evaluation group consists of three evaluations: EXPALL (10950), EXPCURR (10951), and EXPGRM (10953), though variable A is not included in EXPGRM since it has not yet been remeasured on any plots. Variable B has now been collected in inventory years 2008 and 2009 on approximately 1,600 plots, for an average expansion factor of 15,000 acres/plot. Variable B belongs to evaluation group 12209, which consists of two evaluations: EXPALL (10920) and EXPCURR (10921), corresponding to the first two evaluations in the first evaluation group, but using only two panels of plots.

In 2010, the last inventory year of our example, Alabama requires three evaluation groups (12010, 12110, 12310). Evaluation group 12010 includes plots from 5 inventory years (2006-2010). It consists of three evaluations: EXPALL (11050), EXPCURR (11051), and EXPGRM (11053). Evaluation group 12110 includes only one evaluation, EXPGRM (11013), which includes approximately 800 plots that have been measured for variable A at two points in time: 2005 and 2010. This works out to an expansion factor of 30,000 acres/plot. Variable B has now been collected in inventory years 2008-2010 on approximately 2,400 plots, for an average expansion factor of 10,000 acres/plot. Variable B belongs to evaluation group 12310, which consists of two evaluations: EXPALL (11030) and EXPCURR (11031), corresponding to the first two evaluations in the first evaluation group, but using only three panels of plots.

It should be apparent from this example that the maximum number of evaluation groups required to handle new variables is equal to the cycle length, which is 5 years in the eastern states and 10 years in the western states. It should also be apparent that these "extra" evaluations are no longer necessary once a new variable has been collected on a full cycle of plots.

REMAINING ISSUES

As was mentioned earlier, the recommended approach will result in FIA's reporting tools computing higher sampling errors of estimates for new variables in the interim period prior to the completion of a complete cycle of data collection, with the sampling errors being especially large for the first panel. There would appear to be two ways of dealing with this issue. One is to permit reporting on new variables immediately upon the collection and compilation of the first panel of plots. In this case, it is recommended that the reporting tools somehow highlight the fact that the estimates are not based upon a full cycle of data, as well as compute the estimates and sampling errors. The other method would be to permit reporting only after a threshold percentage of plots (e.g. 40 percent or 60 percent) are collected for a new variable. This would somewhat reduce the initial sampling errors, but the reporting tools should once again highlight that fact.

LITERATURE CITED

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