THE NEW SOUTHERN FIA DATA COMPILATION SYSTEM[1]

V. Clark Baldwin, Jr., and Larry Royer[2]

Abstract—In general, the major national Forest Inventory and Analysis annual inventory emphasis has been on data-base design and not on data processing and calculation of various new attributes. Two key programming techniques required for efficient data processing are indexing and modularization. The Southern Research Station Compilation System utilizes modular and indexing techniques applied with standard Oracle tools. We present the unit's approach and describe the challenges encountered as a guide for others embarking on the same complex computational journey.

INTRODUCTION

The new southern Forest Inventory and Analysis (FIA) Inventory Compilation System includes data processing activities from field collection to processed data output. For convenience, the process is divided into three phases: (1) data collection, (2) calculation of derived data, and (3) output of the compiled data. Although relatively straightforward and manageable, the computing process is complex because of the changes in sampling design, addition or redefinition of variables measured, and new or additional outputs required in the shift from periodic to annual inventories.

The complexities begin at the field collection phase, where the data recorder software must be capable of managing input from various sampling scenarios. The first scenario is the remeasurement of the periodic/prism sample with the initial inventory of the annual/mapped sample overlaid on the prism/periodic sample (USDA 1967, 1998). The second scenario is the remeasurement of the annual/mapped sample (USDA 1998). The third scenario is the remeasurement of a fixed-area sample (e.g., 1998 Kentucky plots) with a new initial annual/mapped sample at a different location (USDA 1999). At the present time, nine States are downloading previous inventory data, collecting new data, and transmitting data. While the States are actively converting data from periodic to annual inventories in different stages and ways, data from no two States are similarly compiled. Sophisticated software downloads and formats data from a variety of sources, such as flat-files (Anonymous 2000) and Oracle™ database tables (Koch and Loney 1997), into a common Microsoft™ (MS) Access format (Anonymous 2000). The field crews operating in the remeasurement of the annual/mapped mode can query the Oracle database tables through any remote Internet connection and build the county's previous inventory data interactively on the personal data recorder. Crews operating in the other modes download preformatted historic data to the personal data recorder. They can transmit and capture data through a dial-up connection to a server in Starkville, MS, which loads the data into a set of Oracle relational production database tables.

The amount of data flowing concurrently from nine States is part of the "tidal wave of data" referred to during this conference. From 28,101 sample plots, 21,161 were submitted via data recorder; and 6,940 were submitted by other means. The number of trees from these plots totaled 605,281. Data flowing so quickly into the system caused a logjam. To accommodate this massive data input, we have designed a compilation system and data-flow method to clear the logjam and ensure efficient data processing.

The Southern Compilation System had to overcome many challenges during system development; these involve algorithm development and programming and required the creation of immediate solutions. The difficulties centered in the areas of (1) computer system and Oracle software performance, (2) input data, (3) change accommodation, (4) area reconciliation, and (5) moving-average estimation.

THE SOUTHERN COMPILATION SYSTEM

Concepts

The data compilation system must process current and previous data for all of the scenarios described above. The Oracle database system accomplishes these complicated tasks using a relational database format. The three major groups of relational tables within our Compilation System are (1) Production Tables, (2) Regional Tables, and (3) National Tables.

Indexing—The Compilation System hierarchy was designed with primary index keys that link the database tables to allow quick and easy access to any element. The primary index key, PIX_ID, grows as the level of the table increases within the hierarchy tree. The primary index key is similar to a serial number and can be used to locate an item in any table, based upon its relation to any other table. Levels of processing are State, county, cycle, panel, plot, and individual tree. The following programming code example illustrates the process:

Production Table Prod_Plot plot level PIX_ID = 1300103141 where State = 13, County = 001, Plot = 031, Cycle = 4, and Panel = 1.

Regional Table Inventory tree level PIX_ID = 13001031413010 where State = 13, County = 001, Plot = 031, Cycle = 4, Panel = 1 , Subplot = 3, and Tree Number = 10.

National Table Tree level PIX_ID = 13001031413010 where State = 13, County = 001, Plot = 031, Cycle = 4, Panel = 1, Subplot = 3, and Tree Number = 10.

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

[2] Research Forester and Mensurationist, USDA Forest Service, Southern Research Station, 200 Weaver Blvd., Asheville, NC 28804, respectively.

We establish an SQL query by joining the three tables, using a substring of the PIX_ID key. The following simple SQL statement can query the attributes of species code, diameter from the Tree table, past diameter at breast height (d.b.h.) from the Inventory table, and remeasurement period from the Plot table:

Select **Tree.spcd, Tree.dia, Inventory.Pastd.b.h., Plot.Remper** from **Tree, Inventory, Plot** where **Tree.pix_id = 13001031413010** and **Inventory.Pix_Id = Tree.Pix_Id** and **Plot.Pix_Id = substr(Tree.Pix_Id,1, 10)**

By using the PIX_ID primary index key, we can reduce paragraphs of Structured Query Language (SQL) code to a statement that can locate individual items, such as one tree. Without the PIX_ID primary index key, the query to select the exact same attributes would be:

Select **Tree.Spcd, Tree.Dia, Inventory.Pastd.b.h., Plot.Remper** From **Tree, Inventory, Plot** Where **(Tree.Statecd = 13** and **Tree.Countycd = 1** And **Tree.Plot = 31** And **Tree.Cycle = 4** And **Tree.Panel = 1** And **Tree.Subplot = 3** And **Tree.Tree = 7)** And **(Inventory.State = Tree.Statecd** And **Inventory.County = Tree.Countycd** And **Inventory.Location = Tree.Plot** And **Inventory.Cycle = Tree.Cycle** And **Inventory.Panel = Tree.Panel** And **Inventory.Point_Number = Tree.Subplot** And **Inventory.Tree_Number = Tree.Tree)** And **(Plot.Statecd = Tree.Statecd** And **Plot.Countycd = Tree.Countycd** And **Plot.Plot = Tree.Plot** And **Plot.Cycle = Tree.Cycle** And **Plot.Panel = Tree.Panel)**

In complexity and performance, the PIX_ID process is less complicated and executes faster than code written without this feature. Each tree has a unique PIX_ID that allows the isolation, tracking, and processing of any individual tree throughout the entire system.

Modules—The Compilation System uses modules and functions of Oracle PL/SQL code (Urman 1996) to break down the complex process scenarios into small tasks. To manage maintenance, debugging, and change, each module or function within the system was kept small and limited to one task. Thus, for example, if a volume equation's coefficients change, only the function for volume needs modification or replacement. For debugging a data problem, the data need only be run through an appropriate module or function for the problem to be isolated. Most modules and functions can be executed at any level of processing and in any order. An individual tree with a data problem may be reprocessed using the module or modules in question without reprocessing the whole data set. However, area data can be processed only at the county level due to the nature of the data. Some possible levels of processing are State, county, unit, cycle, panel, plot, and individual tree. In this semi-automated design mode, as a plot clears the internal edit, processing modules at the plot level begin for all of the data on the plot. When the last plot of a county clears the internal edit, the area processing modules are triggered to process the county data. This semi-automated mode provides that all of the data for a processing level will be complete when the last plot clears the internal edit.

The Compilation System modules are divided into four groups: (1) Loader, (2) Stocking, (3) Volume, and (4) Area (fig. 1). Loader modules were designed as a dynamic front-end engine to translate, format, and populate data into national and regional tables. When the data structures, definitions, and variables change within the input data received from the field, the Loader modules can accommodate the changes without affecting the other more static module groups. The National Field Manual will require major system design changes in the Loader modules, but the other modules will need only minor or no modifications. The Stocking group consists of modules that calculate trees per acre, stocking, forest type, and stand size, which use national algorithms. The Volume group modules calculate total cubic foot volume; cubic volume of the sawlog section; board foot volume; growth, removals, and mortality; and weight. The Area group modules calculate forest area, area factors, and remeasurement factors.

The Stocking group consists of four major modules: (1) trees per acre, (2) stocking, (3) forest type, and (4) stand size. The trees per-acre modules calculate both prism plot sample trees per acre and mapped/annual plot trees per acre. Two methods calculate prism trees per acre depending upon the size of the trees. For trees 1.0 to 4.9 inches in d.b.h.:

> Trees per acre = 300 / number of measured sample points.

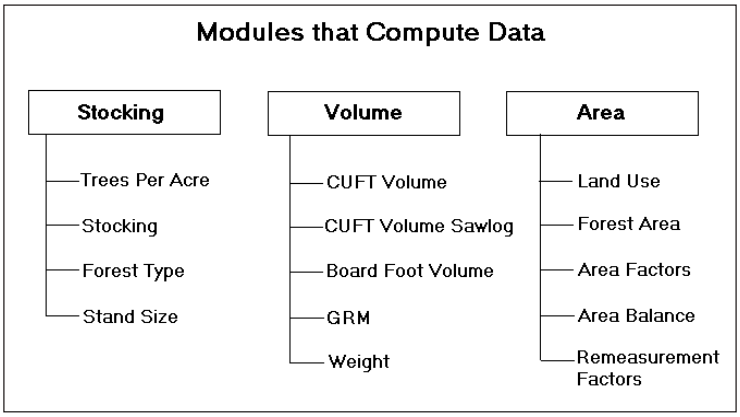

Figure 1—Computation modules of the Southern FIA Compilation System.

For trees = to 5.0 in. d.b.h.,

Trees per acre = 6,875.49354 / [(number of measured sample points) $d.b.h.^2$].

Two methods calculate the annual/mapped trees per acre for trees sampled in a microplot or subplot. For trees within a microplot:

Trees per acre = 43,560 / [(number of microplot points / 100) 581.07].

For trees within a subplot:

Trees per acre = 43,560 / [(number of subplot points / 100) 7,238.23].

A national team of scientists developed algorithms for the Stocking, Forest Type, and Stand Size modules in the Stocking group.

The Volume group consists of five major modules: (1) Cubic-Foot, (2) Sawlog Portion, (3) Board-Foot, (4) Growth, Removals, and Mortality (GRM), and (5) Weight. The volume and weight modules use standard volume equations generally of the form:

Volume or weight = Coeff A + Coeff B ($d.b.h.^2$ Height).

However, any acceptable equations may be used to calculate the tree volumes and weights.

The GRM module is more complex because it must deal with missing items such as diameter and height of cut and dead trees. The general model for GRM growth values is

Growth = (current volume - previous volume) / remeasurement period.

A regression model computes any missing d.b.h.

Predicted Current d.b.h. = (Coeff A) Former Measured d.b.h. $^{(\text{Coeff B})\text{EXP}(\text{Coeff C})\text{ Former Measured d.b.h.}}$.

This produces values for deriving annual radial increments for the area sampled.

By adjusting a height equation for site differences, we can predict the height of a cut or mortality tree from the predicted diameter. First, we predict the current height from equations of the form:

Predicted Current Height = Coeff A + Coeff B $[\text{Log10}(\text{Predicted Current d.b.h.})]^{1/2}$.

Next, we predict a Former Height using equations developed from the same model, with Measured Former D.B.H. replacing Predicted Current D.B.H. We then determine a harmonic proportion by:

Proportion = Measured Former Height / Predicted Former Height.

Finally, we predict Current Height as:

Predicted Current Height = Predicted Current Height (Proportion).

The resulting height is a function of the original height of the tree as well as the diameter. This procedure reflects the influence of the tree site on the height prediction. After calculating a tree's missing variables we can then estimate the growth of removals or mortality. The trees are theoretically grown forward or shrunk backward by the appropriate number of years of growth.

The Area group consists of five major modules: (1) Land Use, (2) Forest Area, (3) Area Factors, (4) Area Balance, and (5) Remeasurement Factors.

The Land Use module loads the photo interpretation information. The Forest Area module interprets aerial photography, field calls, and intensification plot samples corrected for forest, nonforest, and water land-use types.

We divide the county forest area, nonforest area, and water area by the number of conditions within each classification to produce an area factor for each sample. The sample area factor is proportioned by the percentage of each land-use classification and assigned to sample conditions.

The Area Balance module adjusts the area factors at the condition level so that the rounded area factors will equal the enumerated acreage of the county. The Remeasurement Factor module calculates an area factor based upon the previous inventory forest acreage and forest sample plots for GRM expansion. The Southern Annual Inventory uses different procedures for calculating area factors (Reams and Van Deusen 1999), because the current inventory sample must be combined with the previous inventory sample.

Challenges

The overall challenge was to design a system that would accommodate the various existing and planned situations. The system has to accommodate (1) multiple State inventories using different procedures to produce a common set of data, (2) a massive flow of field data, (3) major changes without substantially affecting the processing of data, and (4) a major sample design change due to transition from a periodic inventory to an annual inventory. The SRS Compilation System can accommodate all of these complexities and more.

One specific major challenge was the calculation of area. The national database structure requires that the Area Factors be applied at the plot level and not the condition level where they are calculated. This was particularly difficult for the southern State inventories because Area Factors are calculated for forest, nonforest, and water areas, especially when a sample contained mixed conditions between forest, nonforest and water. Two examples highlight area calculation challenges:

Example 1—Table 1 illustrates an Area Factor situation at the plot level. The first two rows demonstrate a problem that arises when individual condition-level Area Factors are combined into a single plot-level Area Factor. This sample plot contains two conditions: one is a nonforest land use of

code 62 and the other a forest land use of code 20 (col. 1). There are separate Area Factors for nonforest and forest at the county level (col. 6), and these factors have different values (4,423.7331 acres and 6,016.7952 acres, respectively). To obtain the proportioned condition Area Factor for nonforest in this sample, we multiply the county nonforest Area Factor (4,423.7331 acres) by the condition proportion of nonforest (.75, col. 2), which equals 3,317.7998 acres (col. 3). We calculate the forest condition Area Factor (1,504.1998 acres) in the same manner. If we sum these individual condition factors with all of the other sample plot condition factors grouped by their respective land uses, they would equal their respective totals of area for the county. However, if we combine the two condition factors for different land uses *based upon different Area Factors* into a single plot-level Area Factor, they will never equal the correct county acreage for the respective land use. The values in Table 1 illustrate this. The single plot-level Area Factor, derived by summing the nonforest condition factor (3,317.7998 acres) with the forest condition factor (1,504.1998 acres) equals 4,821.9996 acres (col. 4). This value is the only Area Factor that the national database tables carry.

To calculate area for the respective land uses within a county based upon its sample plots, we multiply the single plot-level Area Factor (col. 4) by the condition proportion. For this example, the procedure produces the following results: the nonforest condition factor would equal the plot Area Factor (4,821.9996 acres) multiplied by the nonforest proportion (.75), which equals 3,616.497 acres (col. 5). The forest Area Factor would equal the plot Area Factor (4,821.9996 acres) multiplied by the forest condition proportion (.25), which equals 1205.4999 (col. 5). But now neither of these two values will sum to the correct acreage for their respective land uses at the county level. The nonforest condition factor should be 3,317.7998 acres, but has been calculated by the plot-level Area Factor method as 3,616.4970 acres. This results in 298.6972 too many acres for this condition. The forest condition factor should be 1,504.1998 acres, but has been calculated by the plot-level Area Factor method as 1,205.4999. This results in 298.6669 too few acres for this condition. Thus, since the county-level Area Factors for differing land uses will always be different values (col. 6), this method does not work when there are different land use conditions. On the other hand, this example also illustrates that when a sample plot has two different conditions that are in the same land use (rows 3 and 4), the plot-level Area Factor (col. 4) process does work because they are both calculated using the same county land-use Area Factor (col. 6).

The lower portion of table 1 demonstrates the solution. We must recalculate the condition proportion (col. 2) for each sample that contains more than one land-use condition (col. 1). We accomplish this recalculation by dividing the calculated land-use condition area factor (col. 3) by the county Area Factor (col. 6) and using that value as the condition proportion (col. 2: .688 and .312 for land uses 62 and 20, respectively) in the database record. We then proportion the condition Area Factors by these new values, resulting in 3,317.5857 and 1,504.4628 acres for Land Use conditions 62 and 20, respectively (col. 5). The sum of these new values (which bring along the influence of their original Land Use factor) is 4,821.9985 acres, a value very close to the column 4 value. When we proportion the plot Area Factors by the recalculated condition proportion (col. 5), the resulting values are also very close to the original condition area Land Use factors (col. 5). Note that since these values are not the exact calculated condition Land Use factors,

Table 1—An example of a potential problem in calculating area factors

Land use	Condition proportion	Condition area factor	Plot area factor	Condition area factor proportioned	County area factor
		Plot level area factor challenge			
62	.75	3,317.7998	4,821.9996	3,616.497	4,423.7331
20	.25	1,504.1998	4,821.9996	1,205.4999	6,016.7952
20	.75	2,691.846	3,591.9996	2,693.8845	3,591.8459
20	.25	897.9615	3,591.846	897.9615	3,591.8459

Solution: Condition Proportion = Condition Area Factor / County Land Use Area Factor

CP62 = 3,317.998 / 4,423.7331 CP20 = 1,504.1998 / 6,016.7952

Land use	Condition proportion	Condition area factor	Plot area factor	Condition area factor proportioned	County area factor
62	.69	3,317.7998	4,821.9996	3,327.1772	4,423.7331
20	.31	1,504.1998	4,821.9996	1,543.0387	6,016.7952

there is a small difference in the total acreage when summed at the county level. Rounding of the recalculated condition proportion makes the difference larger.

Example 2—Another challenge has been the calculation of an Area Factor under the moving average estimation approach (Reams and Van Deusen 1999) for the annual forest inventory system. This concept combines the current panel completed with the rest of the sample population. In other words, if panel 1 contains 20 percent of the sample plots just inventoried, and the remaining 80 percent of a State's plots were inventoried during the last survey cycle, to produce statewide estimates we must combine the forest area calculation using the 20-percent sample with the older 80-percent sample. Two other complications offer challenges. There may be a transition from an old to a new photo interpretation methodology, and sometimes the previous compilation methodology used data in a flat-file format, whereas the current data resides within a relational database format.

First, we collected and reformatted all of the flat-file data and loaded it into the current database structure. We calculated a new estimate of forest area and computed a new set of Area Factors for the entire sample. Then another problem surfaced. The 20-percent current sample population level volumes had been calculated using current tree data and current Area Factors. The 80-percent previous sample population-level volumes had been calculated using previous data and previous Area Factors. To put the entire sample into the same context, we had to calculate the 80-percent sample population level-volumes using the current set of Area Factors.

To accomplish that task, we reduced plot volumes to individual tree-level volumes, recalculated trees-per-acre for each tree, and then recalculated population-level volumes using the current Area Factors. Either of two alternative procedures could be utilized: (1) reformat the 20-percent current sample into a flat-file format and combine those data with the 80-percent previous data, recalculating the population volumes; or (2) reformat the previous 80-percent sample into the National database format and then combine the samples. We selected the first method, reasoning that tested and trusted table-building software could build very accurate tables based on the combined data. If there were discrepancies, that method would point to the compilation procedures and not the table-building procedures, thus quickly revealing any existing problems.

Unfortunately, there were discrepancies. As is common in developmental work, thorough testing of the compilation system output revealed that the first procedure did not produce acceptable results. It, thus, became necessary to pursue development using the second procedure.

CONCLUSION

The indexing and modularization techniques are two key procedures in the new Southern FIA Compilation System that make complex compilation situations manageable. The development examples presented, showing a success and an initial failure, represent just a few of the many challenges we encountered with the new system. So far, the life cycle of the system's processing phase has required 6 months of planning and design, 9 months of initial application programming by one application developer, and 7 months of testing, debugging, and modifications. Unit personnel have concentrated their effort for more than a year on algorithm development, programming, testing, and documentation for all modules, and on computer system design, development, implementation, and maintenance. A conservative estimate of time already spent on this project is about 12,000 person-hours. Final development, testing, and debugging are currently in progress using actual data in a production-type mode of operation. The system will be operational in 2001.

LITERATURE CITED

Anon. 2000. Computer desktop encyclopedia [CD-ROM]. Vers. 14.1. Point Pleasant, PA: The Computer Language Company, Inc., December 2000.

Arner, S.L.; Woudenberg, S.; Water, S. [and others]. National algorithms for determining stocking class, stand size class, and forest type for forest inventory and analysis plots. [Documentation available from the Northeastern Research Station, Forest Inventory and Analysis, 11 Campus Blvd., Newton Square, PA 19073].

Gildin, R. 2000. October 17, 2000. Keynote presentation presented at the second annual forest inventory and analysis symposium; 2000 October 17–18; Salt Lake City, UT. Unpublished presentation. On file with: R. Guldin, Forest Service, U.S. Department of Agriculture, Sidney R. Yates Federal Building, 201 14th Street, SW at Independence Ave., SW, 1 NW Yates, Washington, DC 20250.

Koch, G.; Loney, K. 1997. Oracle 8, the complete reference. Berkely, CA: Osborne McGraw-Hill: 9.

Kreines, D.C. 2000. Oracle SQL, the essential reference. 1st ed. Sebastopol, CA: O'Reilly and Associates.

Reams, G.A.; Van Deusen, P.C. 1999. The southern annual forest inventory system. Journal of Agricultural, Biological, and Environmental Statistics. 4(3): 108–122.

Urman, S. 1996. Oracle PL/SQL programming. Berkeley, CA: Osborne McGraw-Hill.

U.S. Department of Agriculture, Forest Service. 1967. Handbook. Washington, DC.

U.S. Department of Agriculture, Forest Service. 1998. Field instructions for the southern forest inventory, A: remeasurement of prism. Version 3. On file with: U.S. Department of Agriculture, Southern Research Station, Forest Inventory and Analysis, 4700 Old Kingston Pike, Knoxville, TN 37919.

U.S. Department of Agriculture, Forest Service. 1998. Field instructions for the southern forest inventory, B: remeasurement of prism. Version 2. On file with: U.S. Department of Agriculture, Southern Research Station, Forest Inventory and Analysis, 4700 Old Kingston Pike, Knoxville, TN 37919.

U.S. Department of Agriculture, Forest Service. 1999. Field instructions for the southern forest inventory, Kentucky—remeasurement of fixed-radius plots. Version 3. On file with: U.S. Department of Agriculture, Southern Research Station, Forest Inventory and Analysis, 4700 Old Kingston Pike, Knoxville, TN 37919.