The Virtual Analyst Program: Automated Data Mining, Error Analysis, and Reporting

W. Keith Moser[1], Mark H. Hansen, Patrick Miles, Barbara Johnson, and Ronald E. McRoberts

Abstract.—The Forest Inventory and Analysis (FIA) program of the U.S. Department of Agriculture Forest Service conducts ongoing comprehensive inventories of the forest resources of the United States. The Northern Region FIA (NFIA) program has three tasks: (1) core reporting function, which produces the annual and 5-year inventory reports; (2) forest health measurements; and (3) scientific analysis of questions and themes that arise from the data.

Annual reports provide updated views of the extent, composition, and change of a State's forests. These reports have a standard format divided into the three broad categories of area, volume, and change. This reporting process also provides important early trend alerts and error-checking functions. Incorporating our understanding of important trends and relationships, and "cautions" to be aware of, the Virtual Analyst program at NFIA seeks to automate the more repetitive functions of producing reports while highlighting any anomalies that might require further investigation.

This paper discusses the program logic and prototype design. We explore the concepts of data mining and the role it plays in the FIA analysis process. Next, we work backward from the Web-based application product to information-generating vehicles that connect to the forest inventory database. Finally, we will discuss the opportunity to expand this report-writing function into a customized, user-defined data query and analysis function.

Introduction

The Forest Inventory and Analysis (FIA) program of the U.S. Department of Agriculture Forest Service conducts comprehensive forest inventories to estimate the area, volume, growth, and removal of forest resources in the United States, in addition to taking measurements on the health and condition of these resources. The program's sampling design has an intensity of one plot per approximately 2,400 ha and is assumed to produce a random, equal probability sample. Four regional FIA programs divide up responsibility of inventorying and analyzing data in the United States and the Northern Region FIA (NFIA) is responsible for 24 States in the Northeast, Upper Midwest and Great Plains sections of the United States. In 15 States of NFIA, the plots in each State are sampled on a 5-year cycle; i.e., each state has 20 percent of its plots inventoried each year, while the remaining states are sampled on a 7-year cycle.

Such a process generates tremendous quantities of data. A portion of the data generated is analyzed and published in annual and more comprehensive 5-year State reports. Although the production of the tabular output is automated, data review, analytical text, and report highlights have typically required a great deal of human input.

Background

An important component of the core reporting function is the production of updated annual reports. The annual reports are the most current estimates of each State's forest resources and frequently are the first alert to emerging trends in forest structure, composition, growth, and mortality. The reports are divided into the three broad categories of area, volume, and change. The annual report is the final phase of a continuous quality control process, evaluating the accuracy of the data

[1] Contact author: W.K. Moser, U.S. Department of Agriculture, Forest Service, Northern Research Station, Forest Inventory and Analysis, 1992 Folwell Avenue, St. Paul, MN 55108. Phone: 651–649–5155. E-mail: wkmoser@fs.fed.us.

from the data collection on the plot to final dissemination of information and knowledge to our customers. The Virtual Analyst (VA) program at NFIA is designed to serve these multiple needs. Incorporating our understanding of important trends and relationships, and our awareness of important alerts (items warranting further investigation), this program automates the more repetitive tasks of report production.

Data-Mining Theory

With the vast quantity of data generated by the inventory process, an efficient and effective knowledge discovery procedure is critical to providing credible and valuable information to our stakeholders. Frawley *et al.* (1991) defined knowledge discovery as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data." These authors further defined knowledge as a "pattern that is interesting (according to a user-imposed interest measure) and certain enough (again according to the user's criteria)." One of the more salient aspects of a continuous forest inventory is its ability to detect change. While knowledge of the total volume or biomass present is useful, an equally important type of information details changes in forest extent, composition, and structure. These change estimates are not only indicators of trends of interest (e.g., greater numbers of trees, less mortality, etc.), but also send a signal to policymakers and land managers that there is the potential for different ecological futures.

Data Mining

Data mining uncovers structure within existing databases. Beyond concerns with data collection issues (Hand *et al.* 2000), one of its principal benefits to NFIA is the opportunity for error checking. One of the unique aspects of data mining is its focus on patterns within the data. While most statisticians are concerned with primary data analysis (data collected with a particular question or set of questions in mind), data miners focus on secondary analysis aimed at finding relationships that are of interest or value to the database owners (Hand *et al.* 2000). In the place of statistical significance, we need to consider more carefully substantive significance: is the effect important (Hand 1998b)? Data mining is not a one-time activity, but rather an interactive process involving the data miner and the domain expert (the analyst assigned to report on the inventory of a particular State), as well as the data (Hand *et al.* 2000).

The two types of structure in data are models and patterns (Hand 1998a). A model is an overall summary of the data or a subset of the data, whereas a pattern is a local structure made up of a subset of the data. For NFIA, it is important to practice data mining that examines both types of data. A model may be a summary of the inventory data for specific attributes, while a pattern may document an interesting facet of the data, such as increased mortality or change in the species composition of the overstory.

While statisticians are concerned with characterizing the likelihood an apparent structure will arise in a data set given no such structure in the underlying process, data miners focus on simply locating the structure. The responsibility for deciding whether the structure has meaning in terms of the underlying process is shifted to the analyst (Hand 1998b).

The analyst is looking for two types of patterns: "real" and "inadequate" patterns. A real pattern is a trend or structure indicating an actual and significant characteristic of the forest. For example, data analysis might discover the occurrence of a tree species not previously known to exist in that location. On the other hand, an inadequate pattern can indicate either an error in data collection or an anomaly in data conversion. An example of an inadequate pattern can occur when a plot is divided into multiple conditions, reflecting its past management or a change in forest type. If the condition "slice" of the plot is small, and there happens to be one large tree in that slice, the expanded basal area per unit area of land might be unnaturally large, which would be a function of the random occurrence of one large tree in a small sample area. At the State level, the first pattern might appear as a critical error, while the second pattern might not matter. The knowledge of the analyst is essential to separating critical from noncritical errors or anomalies. Patterns that can be explained are more likely to be real and are often obvious in retrospect. Many unexpected patterns discovered in a data set during data mining, however, will be attributable to data inadequacy (Hand *et al.* 2000).

FIA Database

By the nature of the two-phase, fixed-area, plot-based sampling conducted by FIA (McRoberts 2005), a relational data model with hierarchical components has been used to create the FIA Database (FIADB) (Alerich *et al.* 2004). FIADB consists of 12 linked database tables. Each database table provides a means of storing data, such as data collected on a sample tree or plot, and computed data attributes. The latest version of the FIADB provides all the data required to estimate resource attributes and associated sampling errors. This structure makes it relatively easy to produce flat files for customers who do not have access to the database. All of the core report tables in the State reports can be produced from these database tables.

In this paper, we present initial work on the VA program that produces not only the core tables for these reports, but also the corresponding figures, maps, and text portions of the reports including highlights and analysis. In addition, we discuss the opportunity to expand this automated report writing into a customized, user-defined data query and analysis function.

Methodology

The VA program provides a single interface with the FIADB that will produce an entire report for a selected area with a few clicks of the mouse. Figure 1 is a flowchart that displays the path from the Oracle database to the final report. The figure portrays a sequence of actions that are currently performed by domain experts (analysts) based on their own experiences and education. Humans, however, are not completely separated from the analytical process. While some of the process will be automated and standardized, analysts still need to check the flagged anomalies and make corrections if needed (fig. 2).

The VA program is written as a Web-based application using various software development programs. These programs will interface with the FIADB using assorted PL/SQL procedures that extract the various resource- and sampling-error estimates presented in the application. All comparisons, logic checks, and computation of highlights are performed using PL/SQL procedures and functions. The user interface in VA allows the user to define the population of interest for the report being generated by using pull-down menus to select the State and

Figure 1.—*Flowchart of information flows in Virtual Analyst.*

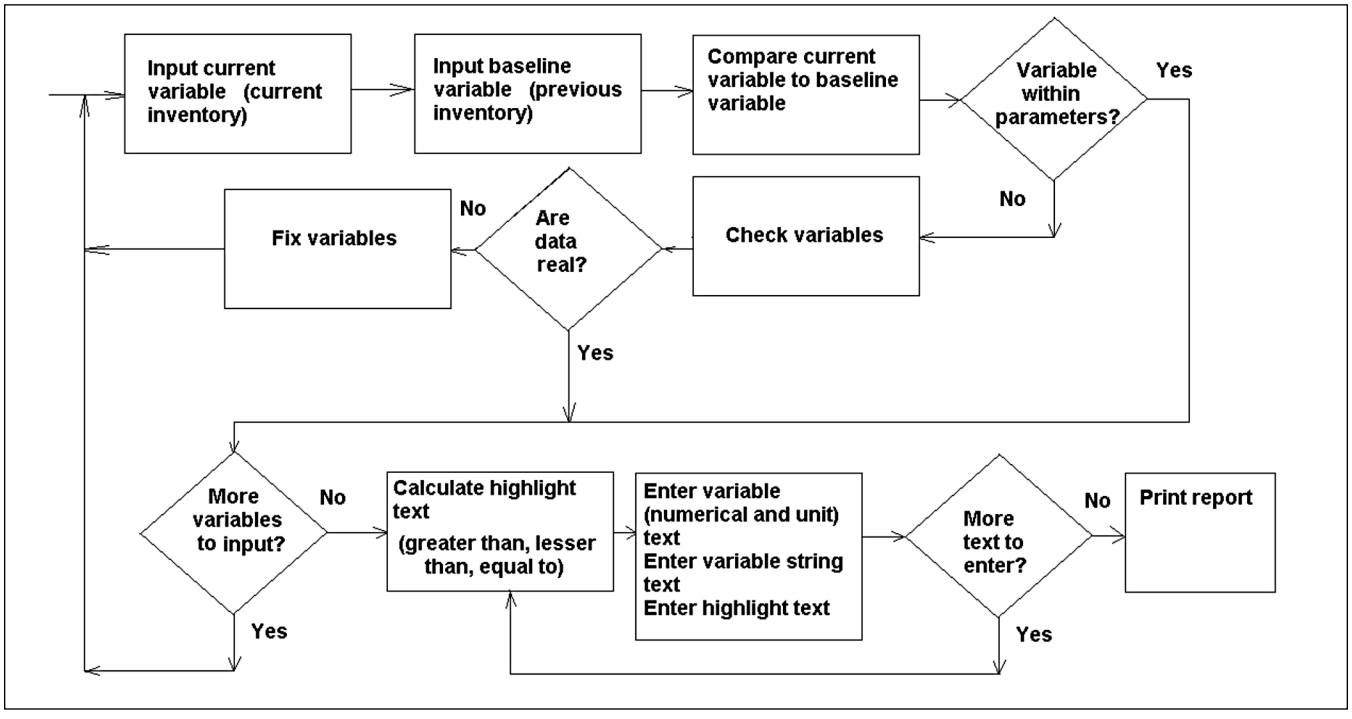
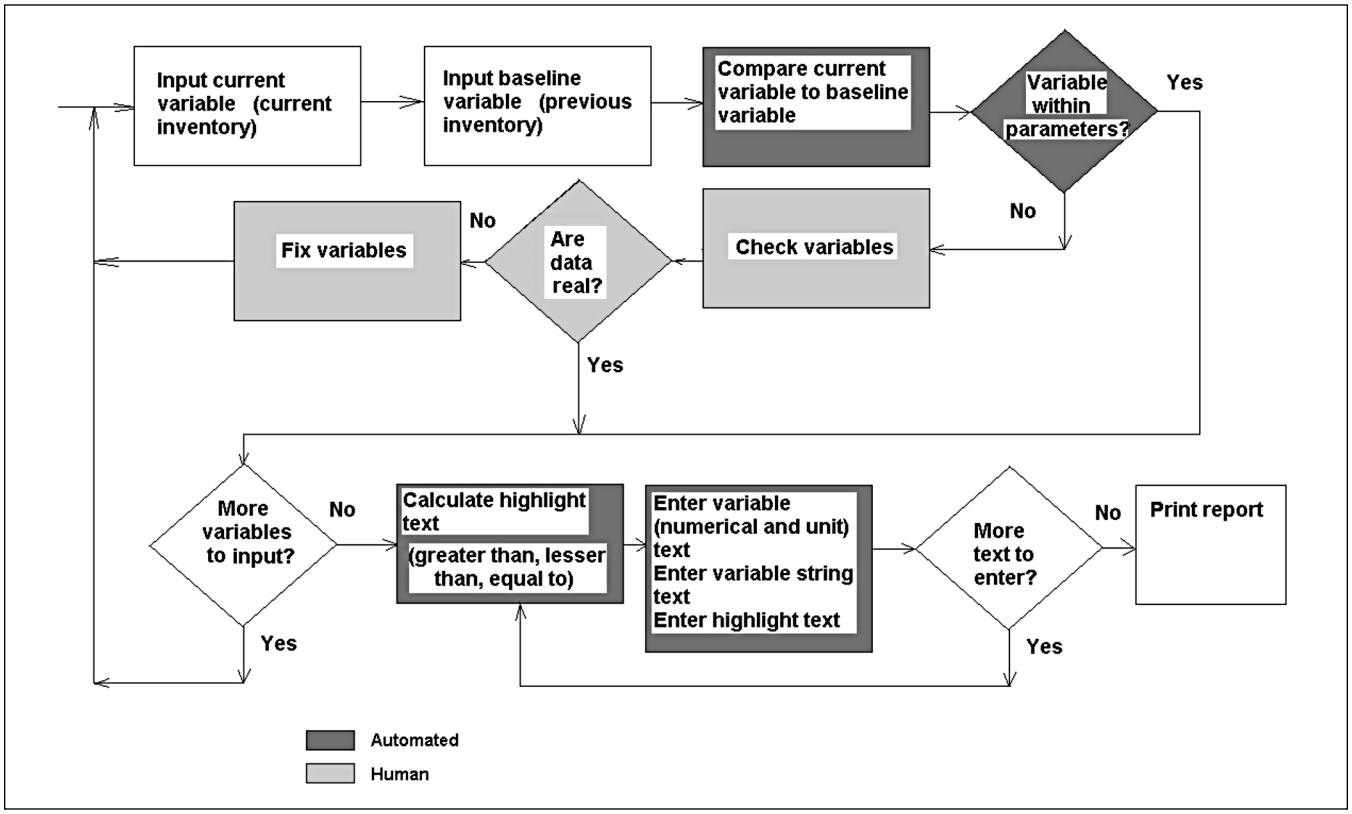


Figure 2.—*Flowchart of Virtual Analyst process, with automated and human-mediated decisions and actions in shaded symbols.*



inventory year. Once the population has been selected, the program executes a series of PL/SQL procedures that generate all the basic report tables for the population (fig. 3). These basic report tables are created as temporary Oracle database tables that exist for the duration of the run and simultaneously display report tables on the Web site and, if requested, as publication-quality PDF files. The VA approach to generating report tables differs from other programs that access FIADB in that catalogs of estimates and sampling errors are created simultaneously in all cases. That is, for every estimate, a corresponding sampling error is computed in a temporary Oracle database table. Other measures of estimation quality, including variance and number of plots where the attribute was observed, are also computed.

Once the report tables are generated for the inventory of interest, identical tables for previous inventories are also computed, using the same PL/SQL procedures. If the data for more than one previous inventory are available for a population, report tables are produced for each of these inventories. These tables are then used to generate charts of the data using .NET charting software.

The estimates and sampling errors in the temporary Oracle database tables can be quickly accessed by simple PL/SQL procedures and functions. These emulate the logic checks and comparisons, and highlight identification procedures typically performed by the analyst (fig. 4). The results of the procedures and functions form the basis of the report text. Output categories from the procedures and functions may be the following:

1. Numerical values (estimates).
2. Units (define the unit of measure for an estimate).
3. Strings (typically describe an estimate).
4. Comments (a special case of a string).
5. Highlights (a special case of a string dependent upon a numerical object).

Numerical values are measured quantities. Units are string objects that define the unit of measure for the values and

Figure 3.—*Flowchart of Virtual Analyst process, with data entry and data checking actions in shaded symbols.*

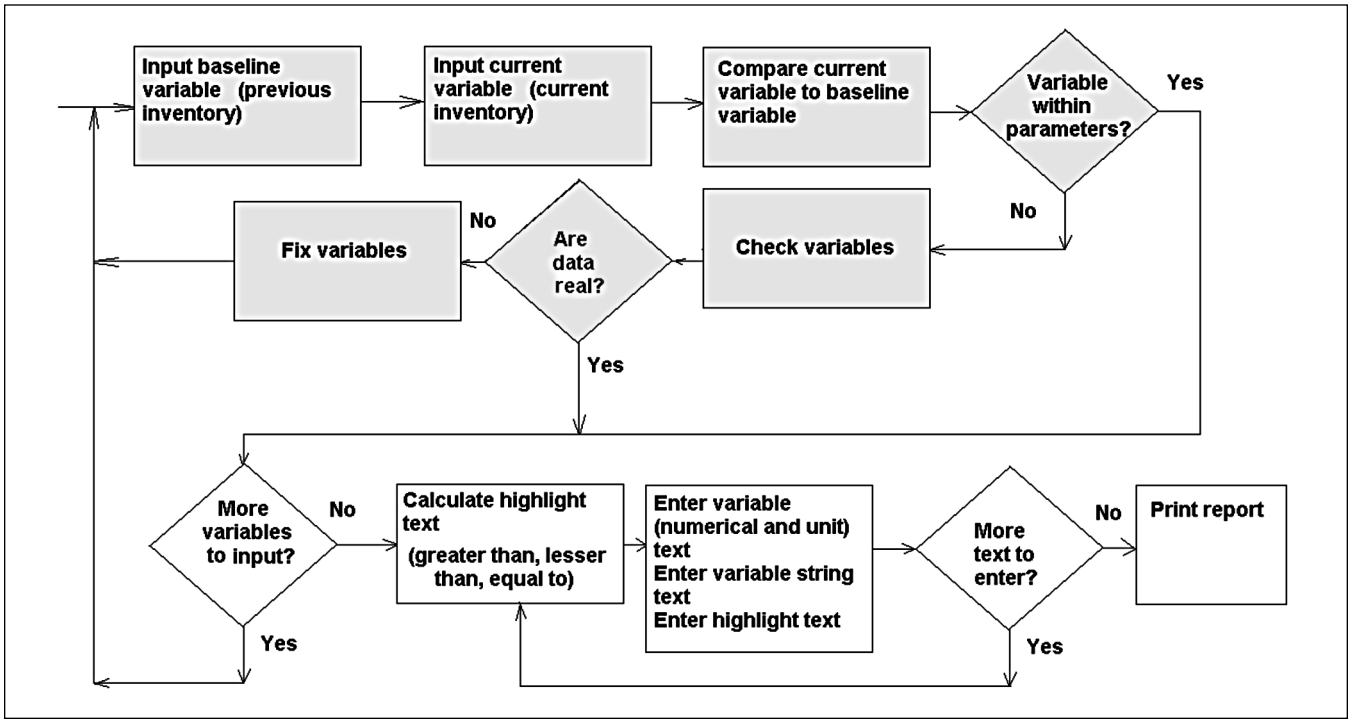
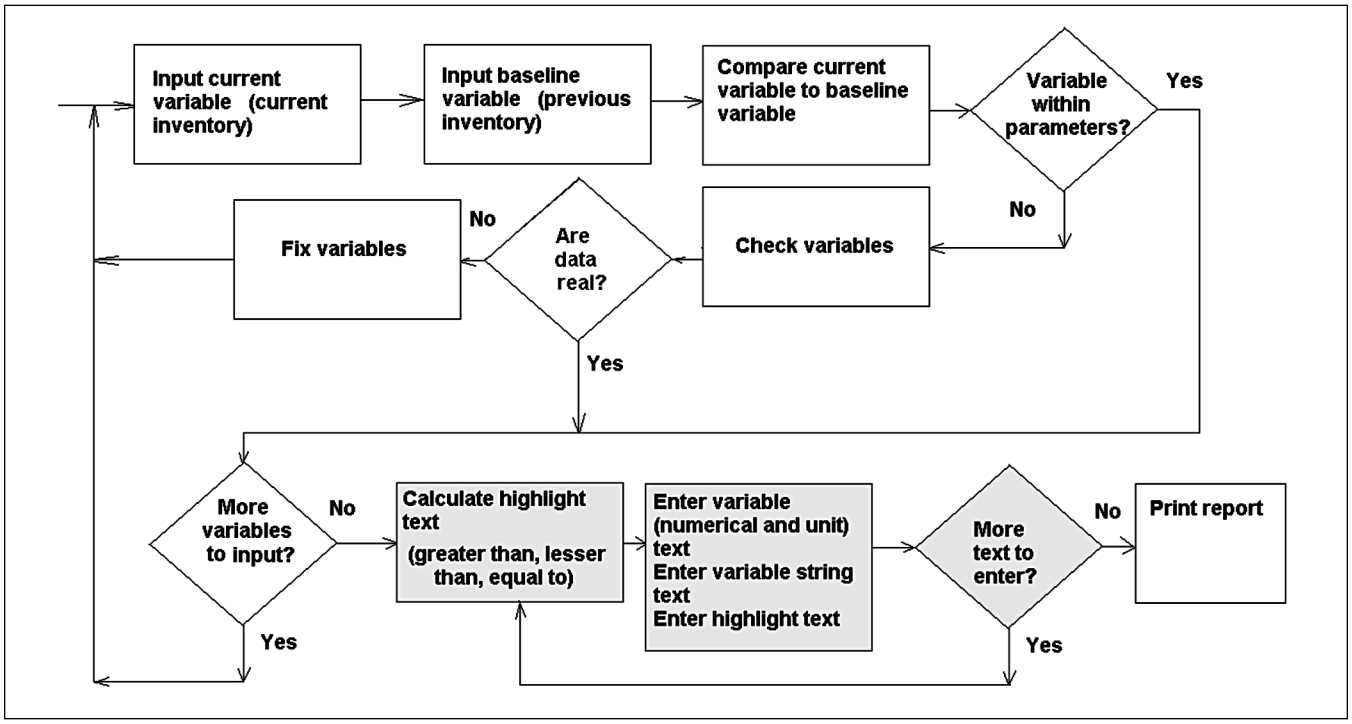


Figure 4.—*Flowchart of Virtual Analyst process, with highlight calculation and text generation actions in shaded symbols.*



are always tied to them. Examples include units of volume, expressed in cubic meters, or area, expressed in hectares or acres. Strings represent categories, such as species or forest type groups. Comments are special types of strings, generally representing a phrase unique to a particular situation. Highlights are terms of precedence or change. Examples include "greater than," "equal to," or "primary."

Calls to procedures and functions that mimic the analytical processes in the data-mining step of the report-writing phase are embedded into the final written report text, such that the resulting text is a complete statement, sentence, or paragraph included in the final document. Some of the functions access a single report table and may simply return the name of the forest type that has the largest estimated area in the most current inventory of all forest types in the population. Other functions may access the same table for multiple inventories, e.g., a procedure that returns the name of the forest type that increased the most in estimated area between the baseline inventory and the current inventory. The complexity and number of tables accessed by these procedures and functions could be increased.

The VA program has the previous inventory tables in the same format as the new inventory, making inventory-to-inventory comparisons relatively easy. Furthermore, with sampling errors for all estimates (including old inventories), we have the potential to conduct statistical tests and avoid highlighting differences that are not statistically significant, or identify small differences that may be statistically significant but that would otherwise go unnoticed.

Results

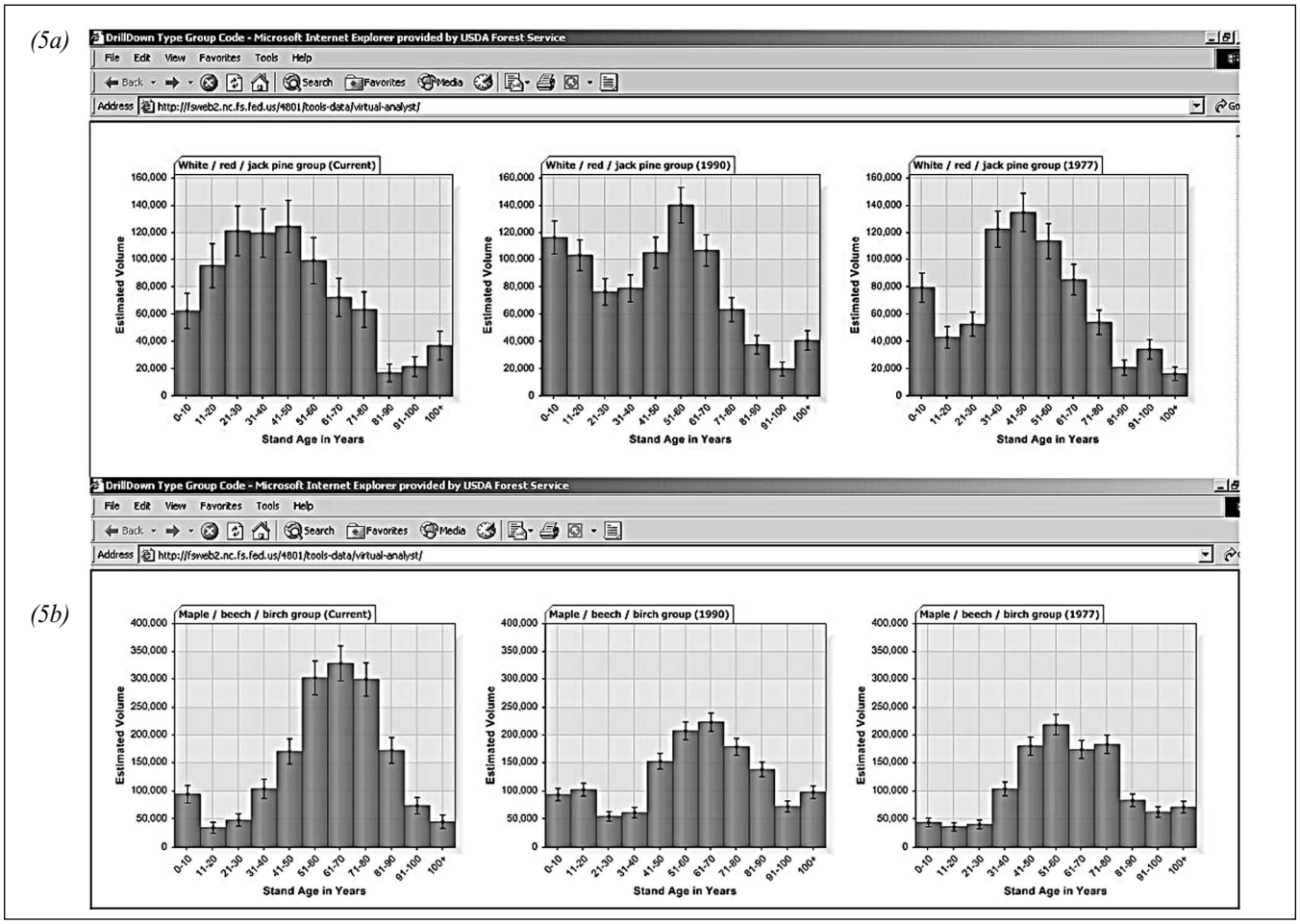
We will limit our discussion to the analysis of some simple report tables, and the associated figures and text portions of the report (highlights and analysis). As an example, we used data from the 1999–2003 Minnesota annual inventory. The VA program generates a basic report table (table 1). Produced at the same time are a matching report table of sampling errors and tables based on the two previous inventories. These tables form the basis of the analysis of forest area by forest type.

Figures 5a and 5b are examples of information produced in association with these tables. The figures show the distribution of timberland area by stand age for the white/red/jack pine and maple/beech/birch forest type groups. The vertical bars at the top of each bar in the histogram show sampling errors for each estimate. These graphics can serve both analytical and illustrative functions. In the pine graph (fig. 5a), the analyst would observe the increased volume in the 11- to 30-year age classes and wish to investigate where these young stands are. The next graph (fig. 5b) provides an example of the illustrative

Table 1.—*Minnesota, 2003 annual estimate, area of timber land by forest type group and stand age class in thousand acres.*

	Stand age class (years)											
Forest type group	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	101+	All ages
White/red/jack pine	62.6	95.9	121.6	120.0	124.9	99.7	72.7	63.6	17.0	21.5	37.1	836.8
Spruce/fir	121.2	131.9	144.3	257.8	436.4	401.0	390.4	362.5	255.9	226.1	490.5	3,218.0
Pinyon/juniper			3.6				4.2	6.5	2.7			16.9
Exotic softwoods			1.6	1.5								3.1
Oak/pine	10.5	22.5	26.1	32.2	33.7	36.4	33.2	17.3	17.8	6.9	3.6	240.1
Oak/hickory	46.2	12.9	16.0	54.0	99.0	178.9	171.0	201.8	145.4	86.7	67.1	1,079.0
Elm/ash/cottonwood	66.5	26.0	50.8	101.7	135.8	171.6	180.0	175.3	95.4	66.2	109.3	1,178.4
Maple/beech/birch	95.2	34.9	48.4	104.7	171.5	303.3	329.4	300.5	173.6	74.5	45.6	1,681.7
Aspen/birch	968.6	722.5	631.2	649.8	874.2	950.5	835.1	427.9	137.0	69.3	31.3	6,297.2
Exotic hardwoods			1.7			2.2						3.8
Nonstocked	204.8											204.8
All groups	1,575.6	1,046.8	1,045.1	1,321.6	1,875.5	2,143.4	2,015.9	1,555.4	844.8	551.4	784.4	14,759.8

Figure 5a and 5b.—*Example of graph output from the Virtual Analyst program. The data refer to the white/red/jack pine forest type group (top, fig. 5a) and the maple/beech/birch forest type group (bottom, fig. 5b) on timber land in Minnesota, across three inventories: 1977, 1990, and 1999–2003.*

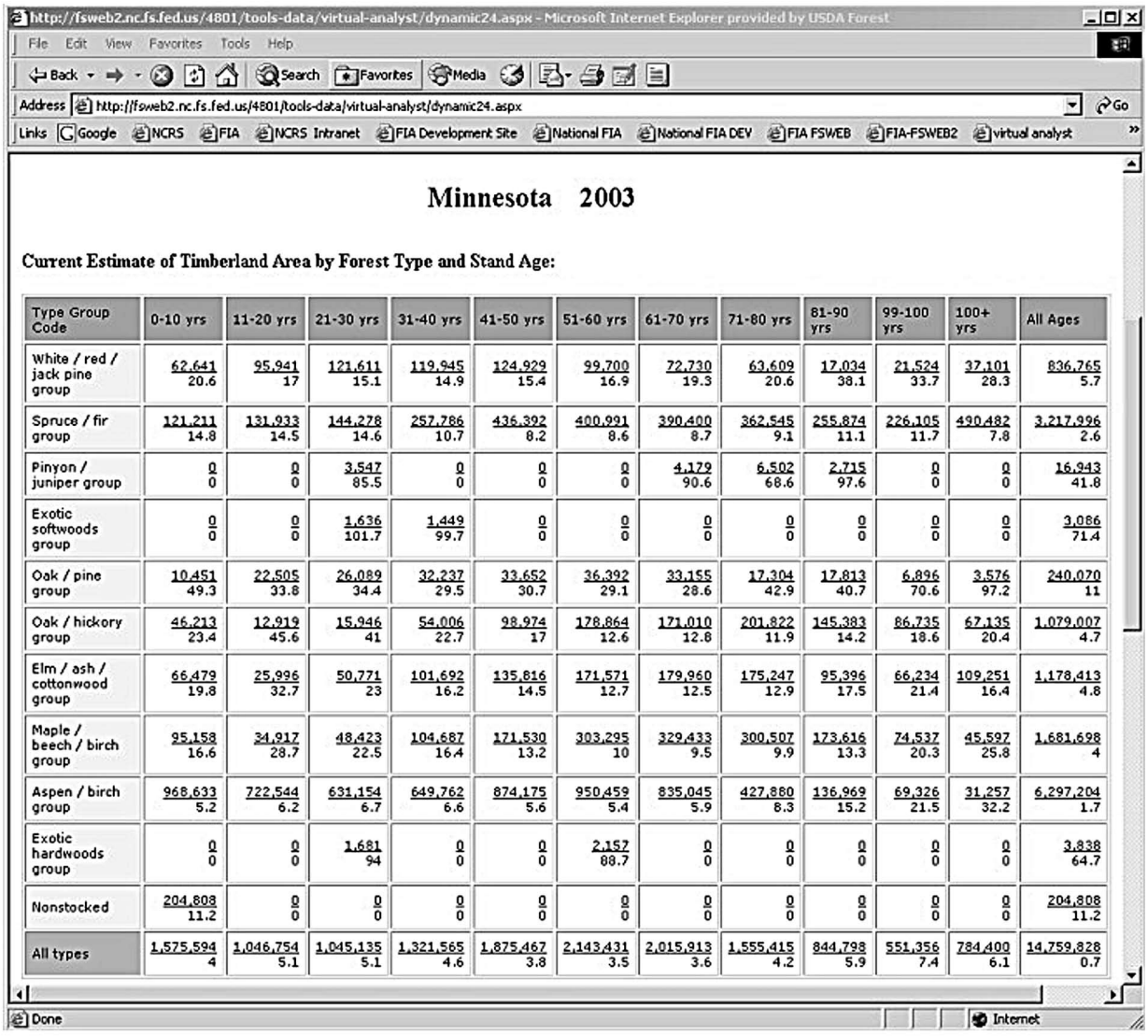


function, where a State resource manager might be interested in seeing the high volumes of wood in the 51- to 80-year age classes. Another example of an illustrative depiction of data is a table of volume, area or other variable of interest, parsed by categories such as age class or forest type (fig. 6). Currently, such tables are generated from the Oracle database into camera-ready PDF files using a separate discrete routine. VA would incorporate final table generation into a seamless data entry–analysis–data display process.

The annual inventory report has traditionally been published as a resource bulletin in the format of several pages of text, including standard definitions and methodology pages, followed by 9 to 12 tables. In the near future, we anticipate that the annual report will be more along the lines of a two-page summary with highlights. Some tables may be published, but most of the data will be available on the Web.

Incorporating this new format, the section below illustrates a hypothetical abstract from the prototype annual report and includes some analysis of the core report table of area by forest type and stand age class and one additional table, volume by species and forest type. The strings and values that can be changed are denoted by output categories in brackets.

Figure 6.—*Prototype tabular output of estimated timberland area by stand age and forest type group generated by the Virtual Analyst program.*



Minnesota 2003

Current Estimate of Timberland Area by Forest Type and Stand Age:

Type Group Code	0-10 yrs	11-20 yrs	21-30 yrs	31-40 yrs	41-50 yrs	51-60 yrs	61-70 yrs	71-80 yrs	81-90 yrs	99-100 yrs	100+ yrs	All Ages
White / red / jack pine group	62,641 20.6	95,941 17	121,611 15.1	119,945 14.9	124,929 15.4	99,700 16.9	72,730 19.3	63,609 20.6	17,034 38.1	21,524 33.7	37,101 28.3	836,765 5.7
Spruce / fir group	121,211 14.8	131,933 14.5	144,278 14.6	257,786 10.7	436,392 8.2	400,991 8.6	390,400 8.7	362,545 9.1	255,874 11.1	226,105 11.7	490,482 7.8	3,217,996 2.6
Pinyon / juniper group	0 0	0 0	3,547 85.5	0 0	0 0	0 0	4,179 90.6	6,502 68.6	2,715 97.6	0 0	0 0	16,943 41.8
Exotic softwoods group	0 0	0 0	1,636 101.7	1,449 99.7	0 0	0 0	0 0	0 0	0 0	0 0	0 0	3,086 71.4
Oak / pine group	10,451 49.3	22,505 33.8	26,089 34.4	32,237 29.5	33,652 30.7	36,392 29.1	33,155 28.6	17,304 42.9	17,813 40.7	6,896 70.6	3,576 97.2	240,070 11
Oak / hickory group	46,213 23.4	12,919 45.6	15,946 41	54,006 22.7	98,974 17	178,864 12.6	171,010 12.8	201,822 11.9	145,383 14.2	86,735 18.6	67,135 20.4	1,079,007 4.7
Elm / ash / cottonwood group	66,479 19.8	25,996 32.7	50,771 23	101,692 16.2	135,816 14.5	171,571 12.7	179,960 12.5	175,247 12.9	95,396 17.5	66,234 21.4	109,251 16.4	1,178,413 4.8
Maple / beech / birch group	95,158 16.6	34,917 28.7	48,423 22.5	104,687 16.4	171,530 13.2	303,295 10	329,433 9.5	300,507 9.9	173,616 13.3	74,537 20.3	45,597 25.8	1,681,698 4
Aspen / birch group	968,633 5.2	722,544 6.2	631,154 6.7	649,762 6.6	874,175 5.6	950,459 5.4	835,045 5.9	427,880 8.3	136,969 15.2	69,326 21.5	31,257 32.2	6,297,204 1.7
Exotic hardwoods group	0 0	0 0	1,681 94	0 0	0 0	2,157 88.7	0 0	0 0	0 0	0 0	0 0	3,838 64.7
Nonstocked	204,808 11.2	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	204,808 11.2
All types	1,575,594 4	1,046,754 5.1	1,045,135 5.1	1,321,565 4.6	1,875,467 3.8	2,143,431 3.5	2,015,913 3.6	1,555,415 4.2	844,798 5.9	551,356 7.4	784,400 6.1	14,759,828 0.7

- Total timberland area is 14.8 [numerical] million acres [units].
- The aspen/birch [string] type is the predominant forest type on the landscape, making up more than 42 [numerical] percent [units] of all timberland.
- Forest types dominated by softwood species make up more than 27 [numerical] percent [units] of the timberland acreage.
- Spruce/fir [string] is the primary [highlight] softwood component by acreage and volume.
- Between 1990 [date] and 1999–2003 [date], the net volume of growing-stock trees on timberland increased [highlight] by 0.9 [numerical] percent [units], from 15.1 [numerical] billion cubic feet [units] to 15.2 [numerical] billion cubic feet [units].

Conclusions

The VA program facilitates the rapid production of annual reports while introducing automated data-mining and error-checking capabilities. Although the program presented here is a prototype, the full version will allow more rapid dissemination of analytical reports to our stakeholders while ensuring the quality of the data contained therein. Future versions of this program could include custom report generation available directly to the stakeholders, allowing fully interactive data analysis and report production. The VA program is just one component of a suite of data analysis/data quality tools, including a sophisticated statistical analysis tool, being developed by the Pacific Northwest FIA program. These tools will enhance resource analysis and improve information quality assurance for the national FIA program.

Literature Cited

Alerich, C.L.; Klevgard, L.; Liff, C.; Miles, P.D. 2004. Forest inventory mapmaker web-application. Ver. 1.7. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station. www.ncrs2.fs.fed.us/4801/fiadb/index.htm.

Frawley, W.J.; Piatetsky-Shapiro, G.; Mathaeus, C.J. 1991. Knowledge discovery in databases: an overview. In: Piatetsky-Shapiro, G.; Frawley, W.J., eds. Knowledge discovery in databases. Cambridge, MA: AAAI/MIT Press: 1-27.

Hand, D.J. 1998a. Data mining: Statistics and more? American Statistician. 52: 112-118.

Hand, D.J. 1998b. Scientific method and statistics. In: Armitage, P.; Colton, T., eds. Encyclopedia of biostatistics. Vol. 5. Chichester, United Kingdom: John Wiley: 3967-3971.

Hand, D.J.; Blunt, G.; Kelly, M.G.; Adams, N.M. 2000. Data mining for fun and profit. Statistical Science. 15(2): 111-131.

McRoberts, R.E. 2005. The enhanced forest inventory and analysis system. In: Bechtold, W.A.; Patterson, P.L., eds. 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: 1-10.

Miles, P.D. 2001. Forest inventory mapmaker users guide. Gen. Tech. Rep. NC-221. St. Paul, MN: U.S. Department of Agriculture, Forest Service, Northern Research Station. 52 p.