Phase 2 and Phase 3 Presentation Grids

Joseph M. McCollum[1] **and Jamie K. Cochran**[2]

Abstract.—Many forest inventory and analysis (FIA) analysts, other researchers, and FIA Spatial Data Services personnel have expressed their desire to use the FIA Phase 2 (P2) and Phase 3 (P3), and Forest Health Monitoring (FHM) grids in presentations and other analytical reports. Such uses have been prohibited due to the necessity of keeping the actual P2, P3, and FHM grids confidential. The Environmental Protection Agency's Environmental Monitoring and Assessment Program (EMAP) grid is often touted as a secure alternative to the FHM grid for legacy FHM plots, but the current FIA P3 sampling frame differs from its FHM predecessor. This article offers an approach involving a weighted bipartite matching algorithm using an alternative to fuzzing and swapping coordinates for FIA P2 and P3 plots to assign plots to EMAP grid points by county and shows preliminary results that keep the actual P2, P3, and FHM grids confidential.

Introduction

In a three-phase sampling design implemented by the Forest Inventory and Analysis (FIA) program, Phase 1 (P1) provides the initial estimate for forest area in double sampling, Phase 2 (P2) is the set of plots from which forest inventory data is collected, and Phase 3 (P3) is a subset of Phase 2 plots (1/16th of P2; details may be found in McCollum and Cochran (2005) from which forest health data is collected. (USDA Forest Service 2005).

For landowner security reasons, FIA data is subject to the Food Security Act of 1985, which states that no one may "disclose such information to the public, unless such information has been transformed into a statistical or aggregate form that does not allow the identification of the person who supplied particular information," (7 U.S.C. 2276). Because FIA plot locations are selected using FIA's hexagon grid sampling array, neither FIA plot location coordinates nor are FIA hexagon locations released to the public. In McCollum and Cochran (2005), for example, several proposed maps showing portions of the P2 and P3 grids were rightfully withheld from the final FIA proceedings publication to comply with this precaution of not releasing hexagon locations. Hence, only maps containing hexagons and nondescript State or county lines were shown, which makes it impossible to link a point on a map to its actual ground location. Since McCollum and Cochran (2005) was presented at the fifth annual Forest Inventory and Analysis Symposium in 2003, FIA has adopted a strategy for releasing plot data, outlined in Lister *et al.* (2005), that is informally known throughout the FIA program as "fuzzing and swapping." It is also referred to by McRoberts *et al.* (2005), as "perturbing and swapping," which involves perturbing the coordinates of the plots so that they are untraceable to any certain location on the ground. In areas where legacy plots are rare, the geometry of the P2 grid is preserved because not more than one plot is spatially proximate to any given grid point. In areas where legacy plots are abundant, the geometry of the P2 grid is not preserved. In neither case is the geometry of the P3 grid preserved. Preserving the geometry of the grid leads to generally equal-sized cells.

The Environmental Protection Agency (EPA) Environmental Monitoring and Assessment Program (EMAP) sampling design incorporates a systematic triangular grid to ensure random selection and appropriate spatial distribution of samples (Overton *et al.* 1990, White *et al.* 1992). EMAP cells are equal in size to the initial FHM cells which were 27 times the size of P2 cells (McRoberts *et al.* 2002). However, when the Forest Health Monitoring program was implemented, the FHM grid was shifted with respect to the EMAP grid.

---

[1] Information Technology Specialist, U.S. Department of Agriculture (USDA) Forest Service, Forest Inventory and Analysis Unit, 4700 Old Kingston Pike, Knoxville, TN 37923.

[2] Supervisory Forester, USDA Forest Service, Forest Inventory and Analysis Unit, 4700 Old Kingston Pike, Knoxville, TN 37923.

When EMAP cells are decomposed by a factor of 27, then cells resembling P2 cells are produced but are offset. In many cases, what the analyst desires is to give the reader an idea of the sample design of FIA and FHM.

A related objective is to summarize FIA plot attributes within hexagon cells and spatially portray these summaries (e.g., choropleth maps of forest proportion, forest volume). The objective of this study is to create a spatially dispersed grid adequate for public presentation purposes that simulates the FIA sample design.

Methods

McCollum *et al.* (2008) show how to decompose the P2 grid to produce a P1 grid used in double sampling for stratification when strata weights are estimated. In similar fashion, we decomposed the EMAP grid by a factor of 27 to produce P2-sized cells. We then used these EMAP-derived grid points to emulate FIA's P2 grid. This process involved creating a threefold decomposition of a hexagon, where it was necessary to collect the center as well as the nodes, and then constructing Thiessen polygons of those points. The resultant cells are hexagons that are one-third the size of the original hexagons. This process was repeated three times to construct hexagons that are 1/27 the size of the original hexagons. Because the EMAP grid is in the public domain, it is not necessary to keep the distance and direction of this shift confidential.

Creating the P3 grid presents some tribulations. At the initial establishment of FIA's three-phase sampling design, 1 FHM plot per 27 FIA P2 plots was present, reflecting the 27-factor enhancement of the FHM grid to create FIA's P2 grid. Thereafter, the former FHM sample intensity was supplanted by a P3 sample intensity of 1 P3 plot per 16 P2 plots. A grid was created by referencing the P3 plots as they were initially assigned by the developer of the P3 plot locations using the P2 hexagon number. Plots with coordinates previously believed to be true later were later discovered to not be inside their intended P2 hexagon. In one FIA unit (Southern Research Station-FIA), P3 plots were not reassigned unless they fell outside their intended P3 hexagons into another P3 hexagon that was already populated or a P3 hexagon that should not be populated.

In States that historically had FHM plots, the resulting ratio ended with 1 P3 plot per 16.2 P2 plots. In the remaining States, the ratio was 1 P3 plot per 16.0 P2 plots. Details may be found in McCollum and Cochran (2005). In the historic FHM States, P2 cells were spatially combined into groups of nine, and five-ninths of the groups were actually filled with a P3 plot. In the remaining States, P2 cells were combined into groups of 16, and each cell was filled with a P3 plot.

The procedure for allocating plots in the new States is more consistent with the Goodchild criteria, cited by Carr *et al.* (1997), Clarke (2002), Kimerling *et al.* (1999), and Song *et al.* (2002). One of the Goodchild criteria is that points be equidistant. Under the rule where the States had previously measured FHM plots. this criterion is not met. For example, there is a short distance from one point to the next in three directions, and a long distance from that same point to the next point in three other directions. Another criterion stated by Goodchild (1994) is that each area contains only one point. Where cells are defined as the historic FHM cells, some cells contain as many as three points while others contain only one and some are empty. Alternatively, if cells are defined as the Thiessen expansion of the existing points, then at least three other Goodchild (1994) criteria are violated: areas are not equal in size or shape and edges are not equal in length. A fourth view is that the network of P3 plot locations is not a grid at all but just a network of suggested plot locations. Because this is a presentation grid, however, and presentation points do not represent actual data points, and because even to this day no standard method exists for imputing data from P3 plots to all P2 plots, we will present one selection rule rather than two, such being a 16-fold composition of the P2 grid.

Because the FHM grid is shifted with respect to the EMAP grid, the FHM-derived FIA grid also is shifted with respect to the EMAP grid. In general, State and county lines are not conformable with hexagons; slightly different numbers of grid points are present at both the P2 and P3 levels, as shown in table 1.

Table 1.—*Southern Research Station State counts of Forest Inventory and Analysis Phase 2 and Phase 3 plots from sampling and presentation grids.*

	Sampling		Presentation		Difference	
	Phase 2	Phase 3	Phase 2	Phase 3	Phase 2	Phase 3
Alabama	5,655	351	5,650	355	5	−4
Arkansas	5,732	361	5,735	361	−3	0
Florida	7,090	446	7,099	441	−9	5
Georgia	6,406	392	6,401	398	5	−6
Kentucky	4,359	276	4,358	273	1	3
Lousiana	5,591	352	5,584	352	7	0
Mississippi	5,221	327	5,223	327	−2	0
North Carolina	5,800	357	5,798	363	2	−6
Oklahoma	7,535	473	7,533	470	2	3
South Carolina	3,453	211	3,445	217	8	−6
Tennesee	4,538	282	4,545	285	−7	−3
Texas	28,955	1,807	28,955	1812	0	−5
Virginia	4,613	283	4,610	283	3	0
Southern Research Station	94,948	5,918	94,936	5,937	12	-19

The presentation grid has already attracted a market. Because perturbed and swapped locations were not yet available for all of Florida, the presentation grid points simulated approximate plot locations rather well in Brown (2007). These presentation grids meet analysts' needs for portraying FIA's P2 and P3 sample design without releasing FIA's actual sampling grids. The next logical step after creating a presentation grid is to assign plots to grid points such that an analyst can portray actual plot intensity and corresponding attributes. This process required the use of the "assignment problem" algorithm, discussed by Edmonds (1965) and Gabow (1976) and then implemented by Rothberg (1985) in the public domain programs MATCH and WMATCH, as well as the "minimum cost network flow problem" discussed by Kennington and Helgason (1980) and implemented by them in the public domain program NETFLO in 1988.

Case Study

To see if we could develop methods for assigning plots to grid points, we started with Georgia because it has the smallest counties of any State in the South, and thus finding a solution would be the most difficult: a higher proportion of plots would cross county boundaries when plots are assigned to presentation grid points that are offset from the original sampling grid points.

The first step involved assigning P3 plots to P3 cells. Desirable features of any perturbing and swapping routine are that plots remain in their counties and that source plots remain close to their target locations. Ideally, the target locations would be ecologically similar as well in terms of land use, forest type, and productivity class (Lister *et al.* 2005). Other than attempting to match census water status, we did not consider these attributes.

Operations research analysts know this problem as the "assignment problem," or the "bipartite (weighted) matching problem." This problem is often described as assigning agents to jobs, and, if the cost of assigning agent i to job j is C_{ij}, what is the optimal assignment?

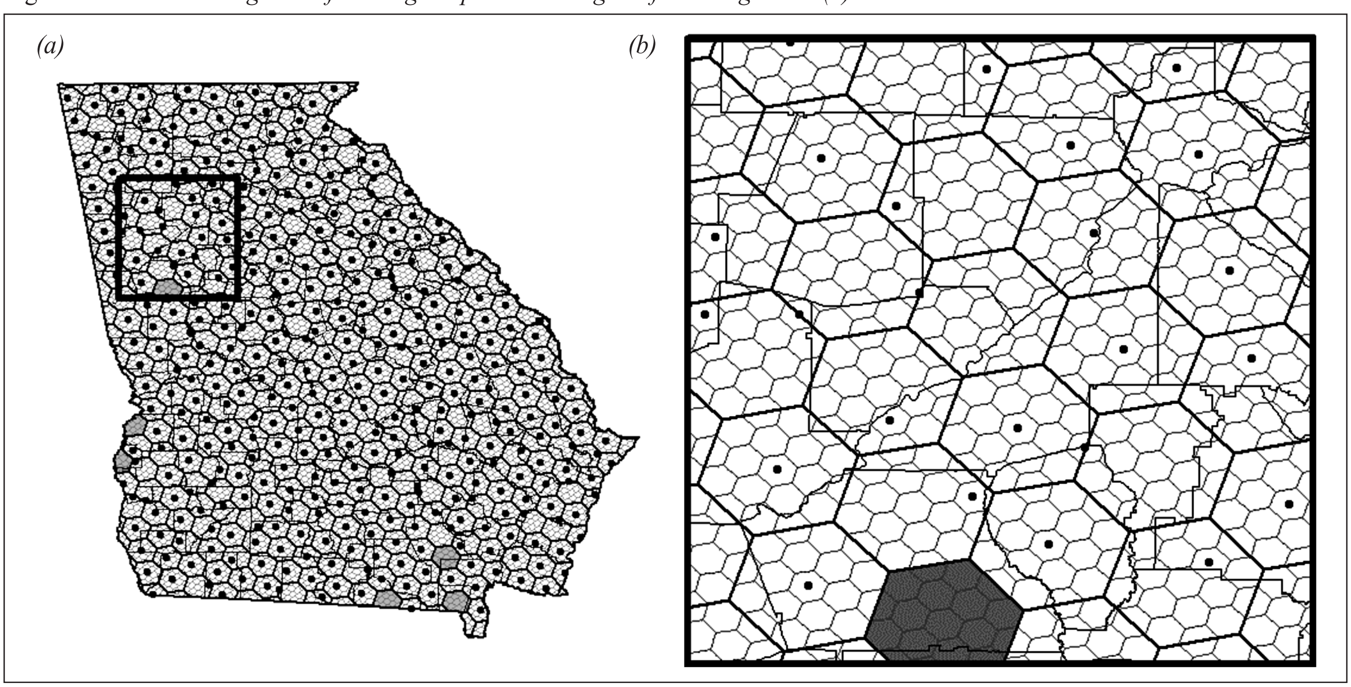
This problem was worked on by Edmonds (1965) and then Gabow (1976) and then operationalized by Rothberg (1985), who wrote C programs called WMATCH for weighted matching and MATCH for unweighted matching. The WMATCH program was used in several stages of our assignment because it was necessary to assign plots to perturbed locations under certain constraints, namely keeping plots in the same county, in unique cells, and reasonably close to their perturbed locations.

Thus, the authors constructed a mathematical graph wherein one type of node was the plots and the other type was the P3 cells. Edges were constructed if the county of the source plot matched the county of the target cell or the county of any remnant polygon within the target cell. The cost was related to the distance between the source plot and the target center, plus a small penalty if census water status did not match, plus a random component designed to prevent unswapping. This graph was input into the WMATCH program, which produced a solution shown in figure 1a, with a detailed view shown in figure 1b. The six P3 cells that could not be matched for Georgia in table 1 are highlighted in figure 1a. In figure 1b, one can see that each P3 cell is composed of 16 P2 cells; actually, 13 whole P2 cells and 6 half cells are present.

The next step involved assigning P2 plots into hexagons that were entirely within one county. Performing this match reduces the dimension of the problem. The results are shown in figure 2a, with a detailed view shown in figure 2b. As in figure 1, colored hexagons are unassigned, and assigned grid points are shown by black dots. Unassigned grid points are shown with empty circles.

A number of unassigned hexes and unassigned plots remain. They could be assigned by trial and error within correct counties, but a better method is available. This method involves consideration of the "minimum cost network flow" problem. This problem involves nodes of supply (where the initial value of the node is the number of available units, recorded as a positive number—one could think of it as the number of agents at this location), nodes of demand (where the initial value of the node is the number of units which are desired, recorded as a negative number—one could think of it as the number of jobs at this location), and possibly intermediate nodes whose initial value is 0. The nodes are connected by edges, which carry a minimum number of units (possibly 0) and a maximum number of units (possibly unbounded), at a specified cost per unit flow. The problem is to transfer units such that the final value of each node is 0. It may be necessary to introduce dummy nodes into the graph. This problem was discussed by Kennington and Helgason (1980), who later wrote a Fortran program called NETFLO (Kennington and Helgason 1988) that was released to the public domain.

Figure 1.—*Phase 3 assignment for using the presentation grids for Georgia and (b) a detailed view.*

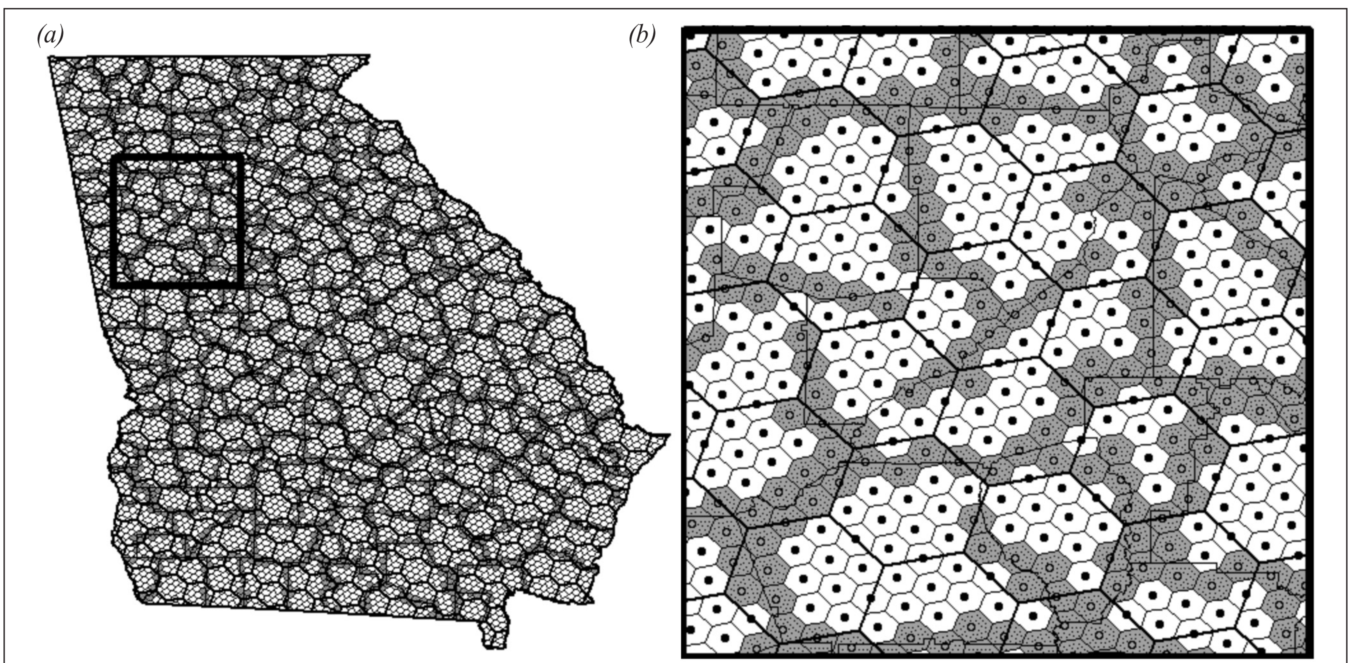


Initially, hexagons belong to the county where their center resides, but a desirable feature of any perturbing and swapping scheme is that the perturbed and swapped location be in the same county as the plot. These features are achieved for plot perturbing and swapping under FIA's current methodology. In terms of the problem, each county was a node, and the initial value was the number of plots not yet assigned minus the number of hexagons not yet assigned. An artificial extra county was introduced that bordered every other county. Its initial value was the number of plots minus the number of target centers; otherwise, edges were constructed between nodes if one county bordered another. The minimum flow was 0, and the maximum flow was the number of hexagon centers along that county border that could be flipped. With the extra county, the maximum flow was 1 if extra plots were present, no more than one could come from any one county, and if defects were present; again, no more than one could be placed in any one county.

The cost of flipping a suggested plot location from a hexagon center to a remnant polygon was inversely related to the size of the remnant polygon plus a severe penalty if the hexagon was shared by three or more counties. The programmatic bookkeeping of flipping a hexagon from county A to county C was much more difficult if it was possible that the hexagon had already been flipped to county B. Thus, such hexagons were not flipped unless it was absolutely necessary, and it was not.

Once it was determined which hexagons should be flipped, it was possible to assign the remaining plots to the remaining plots with WMATCH. This step left five plots unassigned. These plots were assigned to hexagons that had already been assigned. The authors remind readers that, in the days of the 100-second rule (Guldin *et al.* 2006), it was common for multiple plots to be assigned to one location; a Geographic Information System analyst would know to jitter the points (add or subtract a small amount to points that were assigned the same coordinate) before actually making a map. In the event that fewer source points than target points are present, temporary plots could be added that would carry no acres or would not contain any trees but would be merely a forest/nonforest call that would most closely match the actual acres surveyed to be forest with the number of acres that appear to be forest on the map. For example, if a county were more than 50 percent forested and it had 99 plots but 100 hexagons, the extra hexagon would be colored forest if the county itself had 49 or fewer forested plots; if not, it would be colored nonforest.

Figure 2.—*First stage of Phase 2 assignment using the presentation grids for (a) Georgia and (b) a detailed view.*



Discussion and Conclusions

We believe that this method simulates the actual sample design rather well, and, in view of the fact that analysts want to explain the hexagon-based sample design, this grid enables that possibility. We further claim a grid such as this one comes closest to enabling the release of microdata while nearly maximizing the uncertainty of plot locations. The circumscribed radius of a P2 hex is 3,040 m (1.89 mi). Thus, the maximum distance between the true location of a P2 plot and its assigned presentation grid point is 3,040 m (1.89 mi) for those P2 plots occurring within the same presentation hexagons as their assigned presentation grid points. Deviations may be even greater for P2 plots assigned to presentation grid points of presentation hexagons for which the P2 plot locations are outside. These deviations—effectively, perturbing distances—are substantially greater than currently allowed under FIA policy: plot locations may be perturbed up to 1 mi, but many perturbed locations are within 0.5 mi of true plot location coordinates. Indeed, target locations are not tied to actual plot locations at all, unlike with the current method. The P3 signal has been filtered out of the target locations.

Because target locations are not tied to source locations, unfuzzing the data becomes much harder. The proposed method provides an alternative that would be technically feasible if FIA policies were to change. Because target locations are not tied to source locations, however, the presentation grid approach may be less useful than existing public plot location coordinates for remote sensing applications and for geospatial analyses that require the association of FIA plots with specific spatial features. On the other hand, each plot is in its own hexagon, and, if the presentation grid is released, cell-based modeling is enabled.

Potentially, we could consider P1 data when assigning plots to cells. Thus, forested plots would be assigned where forest is most likely to occur.

This method leads to equally sized cells. One could consider a weighted Thiessen method, with the weights based on the current expansion factor (Miles *et al.* 2001) of the plot. Such a method would not guarantee that the polygon area would be equal to the expansion factor. One can imagine trying to construct a jigsaw puzzle of the State such that the size of each piece is equal to the expansion factor of each plot, and the perturbed plot location is on the piece. Ideally, the perturbed plot location should not change over time, although the expansion factors will. The authors submit that maximally dispersed points would allow for maximum flexibility in constructing that puzzle.

Literature Cited

Brown, M.J. 2007. Florida's forests–2005 update. Asheville, NC: U.S. Department of Agriculture Resour. Bull. SRS-118. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 39 p.

Carr, D.B.; Kahn, R.; Sahr, K.; Olsen, A.R. 1997. ISEA discrete global grids. Statistical computing & graphics newsletter. 8(2): 31-39.

Clarke, K.C. 2002. Criteria and measures for the comparison of global geocoding systems. In: Goodchild, M.; Kimerling, A.J., eds. Discrete global grids. Web book. Santa Barbara, CA: University of California, Santa Barbara, National Center for Geographic Information & Analysis. http://www.ncgia.ucsb.edu/globalgrids-book. (22 August 2006).

Edmonds, J. 1965. Paths, trees, and flowers. Canadian Journal of Mathematics. 17: 449-467.

Gabow, H.N. 1976. An efficient implementation of Edmonds' algorithm for maximum matching on graphs. Journal of the ACM. 23(2): 221-234.

Guldin, R.W.; King, S.L.; Scott, C.T. 2006. Vision for the future of FIA: paean to progress, possibilities, and partners. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C.; McWilliams, W.H., eds. Proceedings, sixth annual forest inventory and analysis symposium. Gen. Tech. Rep. WO-70. Washington, DC: U.S. Department of Agriculture, Forest Service. 126 p.

Kennington, J.L.; Helgason, R.V. 1980. Algorithms for network programming. New York: John Wiley. 291 p.

Kennington, J.L.; Helgason, R.V. 1988. The NETFLO program for solving minimum cost network flow problem. ftp://dimacs.rutgers.edu/pub/netflow/mincost/solver-1. (1 November 2006).

Kimerling, A.J.; Sahr, K.; White, D.; Song, L. 1999. Comparing geometrical properties of global grids. Cartography and Geographic Information Systems. 26(4): 271-288.

Lister, A.; Scott, C.; King, S. *et al.* 2005. Strategies for preserving owner privacy in the national information management system of the USDA Forest Service's Forest Inventory and Analysis unit. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C.; McWilliams, W.H.; Cieszewski, C.J., eds. Proceedings, fourth annual forest inventory and analysis symposium. Gen. Tech. Rep. NC-252. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 163-166.

McCollum, J.M.; Cochran, J.K. 2005. Forest inventory and analysis and forest health monitoring: piecing the quilt. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C.; McWilliams, W.H., eds. Proceedings, fifth annual forest inventory and analysis symposium. Gen. Tech. Rep. WO-69. Washington, DC: U.S. Department of Agriculture, Forest Service. 222 p.

McCollum, J.M.; Cochran, J.K.; Rose, A. 2008. Discrete global grid for photointerpretation. Res. Pap. SRS-44. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 7 p.

McRoberts, R.E.; Holden, G.R.; Nelson, M.D. *et al.* 2005. Estimating and circumventing the effects of perturbing and swapping inventory plot locations. Journal of Forestry. 103(6): 275-279.

McRoberts, R.E.; Schmidt, T.L.; O'Neill, K.P. *et al.* 2002. Forest inventories in the United States. Environmetrics newsletter. 8(1): 14-17.

Miles, P.D.; Brand, G.J.; Alerich, C.L. *et al.* 2001. The forest inventory and analysis database: database description and users manual, version 1.0. Gen. Tech. Rep. NC-218. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station. 130 p.

Overton, W.S.; White, D.; Stevens, D.L. 1990. Design report for EMAP. EPA/600/3-91/053. Corvallis, OR: Environmental Research Laboratory. 52 p.

Rothberg, E. 1985. The WMATCH program for finding a maximum-weight matching for undirected graphs. ftp://dimacs.rutgers.edu/pub/netflow/matching/weighted/solver-1. (1 September 2006).

Song, L.; Kimerling, A.J.; Sahr, K. 2002. Developing an equal area global grid by small circle subdivision. In: Goodchild, M.; Kimerling, A.J., eds. Discrete global grids. Web book. Santa Barbara, CA: University of California, Santa Barbara, National Center for Geographic Information & Analysis. http://www.ncgia.ucsb.edu/globalgrids-book. (22 August 2006).

U.S. Department of Agriculture (USDA) Forest Service. 2005. Program structure for forest inventory and monitoring. FIA fact sheet series. http://www.fia.fs.fed.us/library/fact-sheets/overview/Pgm_Structure_FS.pdf. (1 December 2006).

White, D.; Kimerling, A.J.; Overton, W.S. 1992. Cartographic and geometric components of a global sampling design for environmental monitoring. Cartography and Geographic Information Systems. 19(1): 5-22.