Geostatistical Estimation Data for the 1997 National Resources Inventory

S. M. Nusser, J. M. Kienzler and W. A. Fuller

Statistical Laboratory

Iowa State University

December 8, 1999

1 Introduction

The purpose of this report is to outline the procedures used to develop geostatistical information for the 1997 National Resources Inventory (NRI). Geostatistical data are incorporated into estimation procedures in such a way that the 1997 NRI database is representative of the geospatial data.

The geostatistical information developed for the 1997 NRI play the role of the "county base data" that were collected in previous NRI surveys. County base data were collected by each state and included total surface area, surface area of federal land, area in large water bodies and large streams, and area in rural roads for each county in the U.S. The county base data were used as surface area control totals in the 1992 estimation. In some cases imputed points were placed in the database to represent data reported in the county base data, but not observed in the sample segments. These points are called *imputed points* or *pseudo points*.

With the advent of new technologies and information sources, a new process for gathering control information was developed for the 1997 NRI. The general objective was to provide a more reliable and efficient source of control information and to improve the geospatial properties of imputed points. The specific objectives in the use of geospatial information were:

1. to develop a digital product using a documented process that clearly shows the basis for control totals,
2. to improve the quality and consistency of control information, and
3. to provide spatial locations for points imputed as part of the estimation process.

New estimation procedures were developed for the 1997 NRI to take advantage of the increased detail available from geostatistical control information (Fuller *et al.* 1999). In particular, geostatistical data were used to identify tracts requiring imputed points, to generate coordinates for the imputed points, to generate county and hydrologic unit control surface areas, and to generate control totals for federal land and large water bodies within counties and hydrologic units.

We begin with introductory material on the role of control information in sample surveys and an overview of geostatistical data needed for 1997 NRI estimation procedures. The source

materials and procedures for creating 1992 and 1997 geostatistical control data are then described. The procedures are presented in more detail in Nusser (1999).

2 Control Data: Basic Concepts

Control data represent known information about a population. Often this information takes the form of totals for subsets of the population of interest. In human population surveys, a typical set of control data is the total number of persons in demographic categories defined by age and gender classes, or the total number of households within geographic subsets of the population. Examples of control totals include the number of males and females in the U.S. population who are 18-30, 31-50, 51-70, and 71+ years of age for a survey of the U.S. population, or the number of households in each county for a survey of California households. Control totals for surveys of the U.S. population or of U.S. households are generally obtained from Census Bureau statistics. For other surveys, control totals may consist of other kinds of units such as surface areas, the number of business establishments, or the number of teachers in a school district. Control data may be derived from a variety of information sources such as digital maps, membership listings, and administrative databases.

The key concept is that control data represent knowledge about the population being studied, whether that population is defined to be the U.S. human population, the land area in a geographic region, or the members in an organization. In a sample survey, a sample weight (i.e., expansion factor) is calculated for each individual sample (or observation) unit, and is a measure of the number of units in the population that the sample unit represents. Control totals are used in constructing sample weights to ensure that survey results are consistent with the known information represented by control totals. In a 1990 survey of Pennsylvania residents, for example, it is desirable for the sum of the sample weights for 26-44 year-old male respondents to be equal to the true population total of 1,714,098 men, 26-44 years of age in Pennsylvania, obtained from 1990 Census statistics. A variety of methods are available to construct weights so that the sum of weights for sample units in a control category is equal to the known total for that category.

3 Overview of 1997 NRI Geostatistical Estimation Information

The NRI sampling universe includes all land in the U.S. and selected territories (Puerto Rico, Virgin Islands, Guam, Northern Marianas). The sampling universe is divided into a set of polygons defined by the intersection of 4-digit hydrologic unit areas (HUAs) and counties (or analogous entities such as parishes). A single HUA $\times$ county polygon is called a HUCCO. For areas outside the coterminous U.S. (Hawaii, Alaska, Puerto Rico, Virgin Islands, Guam, and the Northern Marianas) some modifications were made in the definition of a HUCCO depending on the digital geospatial data available for the area.

Geostatistical information is obtained for each HUCCO in the NRI universe. This information includes surface area totals for federal land and large water within the HUCCO for 1992 and 1997. The NRI definition of large water is that used by the Census Bureau through the 1980 Census, and consists of water bodies $\geq$ 40 acres or streams $\geq$ 660 feet wide at normal pool, as defined in the 1997 NRI instruction manual (U.S. Department of Agriculture 1997). In addition, 1997 NRI estimation procedures are designed to use known location information for

large water and federal land within HUCCOs to more accurately represent the spatial configuration of large water and federal land in the 1997 NRI database.

The HUCCO surface areas used for the 1997 NRI control totals are:

1. the area of each HUCCO,
2. the area of large water within each HUCCO for 1992 and 1997, and
3. the area of federal land within HUCCOs for 1992 and 1997 (large water is not part of federal land).

The large water areas are further segmented into types of water, as described in a later section.

The geostatistical control data for the 1997 NRI are derived from geospatial layers that depict location and extent of large water and federal land for subunits of the NRI sampling universe. Only the 1992 and 1997 surface areas are obtained from the geospatial layers described in this report. Surface areas for 1982 and 1987 are the acres reported in the 1992 NRI adjusted for difference between the 1992 control acres used in 1992 and the 1992 control acres obtained from the 1997 geospatial data (Fuller *et al*. 1999).

The locations of the NRI primary sampling units (PSUs) were digitized by the National Resources Conservation Service (NRCS). This permitted comparison of the location of the primary sampling units with the location of federal tracts and of water bodies. The location information abstracted from the geospatial layers consists of a list of PSUs whose boundaries intersect with the boundary of each large water body and a list of PSUs that overlap with each federal land polygon. The centroids of each federal and water polygon is computed from the geospatial layer. These data are used during estimation and imputation to associate sample information with large water tracts and federal tracts. Hence, the information in the NRI database has the geospatial coordinates from the geospatial layer.

There are several other types of control totals used for estimation in the 1997 NRI that will not be discussed in this report. For example, other control totals include Conservation Reserve Program (CRP) acres by groups of sign-up periods within states, and previously published estimates of surface area for general categories of land cover/use in previous years for each state (e.g., 1992 areas for land cover/use categories published in the 1992 NRI). Geospatial layers derived from Census Topologically Integrated Geographic Encoding and Referencing System (TIGER) files have also been developed to provide auxiliary information on roads and urban areas. These data are not used directly in weighting, but are used as regression variables in small area models for rural transportation and built-up areas. See Fuller *et al.* (1999).

In the remaining sections, information is provided on geostatistical data for HUCCOs, large water, and federal land.

4 Geostatistical Data for Counties, Hydrologic Units, and Primary Sampling Units

4.1 County Boundaries and the U.S. Shoreline

Census TIGER digital line files were obtained from the U.S. Bureau of the Census (see www.census.gov/geo/www/tiger). These files represent 1:100,000 scale maps that delineate county boundaries and water bodies/streams, as well as roads and other land features. This layer

was released in 1995 and represents the most current information available for political subdivisions of the United States.

The TIGER files define an official county boundary layer for internal boundaries that forms the basis for creating the HUCCOs. The remaining boundaries of counties are shoreline boundaries. The National Oceanic and Atmospheric Administration (NOAA) shoreline formed the basis for the shoreline and was modified to match NRI land universe definitions. The shoreline excludes water areas extending beyond the shoreline of the U.S. coast for counties bordering the oceans and Great Lakes. Because land area associated with islands is included in the NRI universe, the shoreline also circumscribes islands belonging to coastal counties. The adjusted shoreline is sometimes referred to as the Lawson shoreline.

4.2 Four-digit Hydrologic Unit Area Boundaries

A 1:250,000 scale U.S. Geological Survey (USGS) digital line file for the 204 four-digit hydrologic unit areas (HUAs) within the coterminous U.S. was obtained from the NRCS National Cartographic and Geospatial Center (NCGC). The U.S. border in this spatial layer differed from that of the Lawson shoreline, in large part due to differences in map scale. Boundaries of four-digit HUAs located on U.S. borders were adjusted to conform to the Lawson shoreline to provide consistency with the NRI-defined U.S. boundaries and to create a fully labeled set of HUAs within the NRI universe.

4.3 Estimation HUCCO Boundaries

A HUCCO data layer was created by combining data layers for county boundaries, four-digit HUA boundaries, and the Lawson shoreline. There are approximately 4,900 HUCCOs in the coterminous U.S., Puerto Rico and the Virgin Islands. To support statistical estimation, 213 HUCCOs that were less than 6,000 acres in size or that contained fewer than six real points were collapsed with neighboring HUCCOs within the county. County boundaries were always preserved in the collapse operation. The combined HUCCOs are called estimation HUCCOs.

4.4 Estimation PSU Boundaries

The locations of PSUs as digitized by NRCS staff were used in assigning PSUs to HUCCOs, associating PSUs with large water bodies and streams, and assigning PSUs to federal land polygons. If a portion of the digitized PSU was located outside of the county from which it was originally selected, the size of the PSU was modified to the correct size by clipping and removing the portion outside the county. The PSU centroid and surface area were determined for each corrected PSU. The data layer containing the boundaries of the corrected PSUs is called the estimation PSU layer. A PSU was assigned to the HUCCO in which the PSU centroid was located. Not all PSUs have been digitized by NRCS. PSUs not digitized were originally given a centroid of zero. The zero was then replaced with the centroid of the HUCCO.

5 Geostatistical Data for Water and Federal Land

5.1 1992 Water Data

The TIGER files were used as the initial draft layer for the spatial extent and location of large water for 1992. Because the delineation of water areas was not the primary focus in developing TIGER files, it was expected that the TIGER large water data would require modification to improve the geospatial representation for 1992 conditions. Water bodies were labeled to designate four subcategories of large water: large streams, large lakes and reservoirs, gulfs and bays, and estuaries (see USDA 1997 for definitions). In combining the TIGER file and the Lawson shoreline a number of water bodies were created as subdivisions of original tracts. The geospatial file was expanded with designations for the created water bodies.

5.2 1992 Federal Land Data

The 1998 release of a digital layer for federal lands at a 1:2,000,000 scale (updated to 1996 conditions) was obtained from the U.S. Geological Survey (USGS). When TIGER information on federal land was consistent with the USGS data, federal boundaries were extracted from the TIGER data. Otherwise, federal boundaries were extracted from the USGS layer. Information on agency ownership was retained, although insufficient resources were available for fully checking agency ownership information in subsequent steps. Because of the coarse scale, the layer required modification, particularly in western areas of the U.S.

In several states, GIS layers were available that were judged by NRCS Inventory Collection and Coordination Sites (ICCS) staff to provide more accurate information on the spatial location of federal land than the USGS federal land layer. These geospatial layers were accepted by ISU in lieu of the USGS federal information, provided that the digital data were of comparable quality and scale to other materials being used to create the 1992 federal land layer.

5.3 Constructing 1992 Water and Federal Land Layers for Estimation

Both the water and federal layer were checked and updated for each of the approximately 3,300 counties, parishes and territories in the NRI universe. Materials used to verify or update the information included county base data collection sheets from the 1992 NRI, USGS quad sheets (mostly at a 1:100,000 scale) and, in western states, quad sheets depicting Bureau of Land Management (BLM) land. Both spatial configuration and surface area were examined for consistency across information sources. In most counties, additional review of the 1992 water and/or federal land data was performed by NRCS staff. Updates were accomplished using USGS 1:24,000 or 1:100,000 digital raster graphics (DRGs) as background reference material for on-screen editing whenever possible. When all counties in a state had been edited and checked, a paper map of the state was produced and used to check for consistency in large water body and stream shapes, and to check for consistent labeling across county boundaries.

5.5 1997 Water and Federal Land Data

Two methods were used to obtain information on changes in large water and federal land that had occurred between 1992 and 1997. First, the 1997 NRI database was used to identify PSUs where the data gatherer had recorded a change from 1992 to 1997 in the area of large water bodies, a change in large streams or a change in federal land ownership. The ICCS was asked to determine whether the change recorded in the PSU had actually occurred. If a true change had taken place, ICCS staff were asked to document the full scope of the change in large water and/or federal land using a USGS quad sheet or similar supporting material. Polygons representing these changes were added to the 1992 layers and labeled to denote changes from 1992 to 1997. If no true change occurred, ICCS staff corrected the PSU data.

The second method of gathering information on changes from 1992 to 1997 was to ask ICCS staff to identify significant changes that had occurred in the state. The main consideration was to identify those changes that would be widely known by residents of the state. The ICCS staff or their designates (e.g., NRCS state office staff) were asked to document these changes on USGS quad sheets or similar materials. This information was used to update the 1992 layer to include polygons that represented gains or losses from 1992 to 1997 in large water and federal land.

5.6 Federal Land with Large Water

The NRI definition of federal land excludes large water by definition. Thus, prior to extracting estimation data from the federal layer, the area associated with any large water polygon that fell within a federal land polygon was removed to define the final federal land polygon.

6 Generating Geostatistical Estimation Data

Geostatistical estimation data were generated for each state from geospatial layers containing the 1992 and 1997 large water, the 1992 and 1997 federal land, the estimation HUCCOs, and the estimation PSUs. The following information is derived from these layers for each state to support statistical estimation.

1. For each estimation HUCCO in the state:
    - the area of the HUCCO, and
    - a list of PSUs whose centroids are located in the HUCCO.
2. For each large water polygon within an estimation HUCCO:
    - the type of water polygon,
    - variables describing the presence and absence of the polygon in 1992 and in 1997,
    - the centroid of the polygon,
    - the surface area of the polygon, and
    - a list of PSUs that intersect with the polygon.

3. For each federal land polygon within an estimation HUCCO:
    - the type of federal land (agency affiliation),
    - variables describing the presence and absence of the polygon in 1992 and in 1997,
    - the centroid of the polygon,
    - the surface area of the polygon, and
    - a list of PSUs that intersect with the polygon.

These data are used as control information for weight calculations, as input for point imputation, and as geospatial locations for the imputed points (Fuller 1999)

Data leading to point imputation and locations for imputed points were used in all states. In some instances, the data from geospatial files required modification to meet NRI definitions of controls. For example, sometimes the spatial layer was too coarse to show private parcels within a tract classified as federally owned (i.e., inholdings). A procedure is given in Fuller *et al.* (1999) that uses the proportion of PSUs containing federal points to obtain an estimate of the acres actually under federal ownership. In some cases no adequate maps were available to identify the boundaries of federal land or large water in the geospatial layer. For those cases, the surface area data derived from other sources were used as the controls in estimation.

The Census Bureau provides official surface area and land area figures for some counties that may differ slightly from the areas extracted from the TIGER file. In interior counties, the NRI control total is the official surface area as defined by the Census Bureau. The control land area for coastal counties with the Lawson shoreline as the county boundary matched the official land area as closely as possible.

7 Additional Considerations

The geospatial layers and the associated metadata furnish documentation for the data on surface area, hydrologic unit area, large water area, and federal land area that will appear in the 1997 NRI final data base. The quality of the final 1997 NRI product depends on all aspects of collection for all sources of data. Hence, a large effort was devoted to developing the underlying geospatial database for use in statistical estimation for the 1997 NRI. In addition to its use in estimation, it is expected that the geospatial layers will be used in conjunction with the 1997 NRI database to produce paper maps, new geospatial materials, and to provide support for future sample selections.

The construction of geospatial layers and statistical estimation are integrally linked. Because the statistical properties of the estimation procedures applied to the 1997 NRI data rely on the fact that the geospatial layers represent good approximations to the geographical features of the landscape, procedures were designed to make the two data sets consistent. Definitions used in the GIS database must conform to long-standing NRI definitions. For example, large water bodies are defined to be bodies ≥ 40 acres in size as has been done in previous NRIs. Also, it is necessary that attribute data in the GIS database agree with data in the NRI point data file. For example, if the point falls on water classified as an estuary, the water body in the GIS database should be classified as an estuary. Finally, the point data file should provide estimates

for all GIS polygons of relevant types. To meet the last requirement, it was necessary to create HUCCO boundaries so that each HUCCO was sufficiently large and contained enough PSUs to support estimation.

Geospatial control data will be needed in future years to create NRI trending databases that contain data consistent with known changes in large water and federal land. The process used to create the 1992-1997 layers provides a foundation for developing new layers. Experiences from the 1997 NRI are expected to suggest improved methods of collecting and integrating geostatistical control information for future survey efforts.

Acknowledgements

We wish to acknowledge the contributions of several collaborators in the development of the geostatistical information base and estimation procedures. The notion that geostatistical data would provide an improvement over past county base data collection procedures was proposed by J. J. Goebel, in the Resources Inventory Division, USDA Natural Resources Conservation Service. G. Lawson and D. M. Thompson of the Natural Resources Inventory and Analysis Institute, USDA Natural Resources Conservation Service, helped initiate the project and were instrumental in developing prototype procedures. M. E. Manion, M. T. Rogers, K.W. Dodd, and R. Dorsch of the Iowa State University Statistical Laboratory contributed to the development, implementation and use of these data. This research was supported by Cooperative Agreements 68-3A75-8-62 and 68-7482-8-351 between Iowa State University and the Natural Resources Conservation Service.

References

Fuller, W. A. 1999. Estimation procedures for the United States National Resources Inventory. 1999 Proceedings of the Survey Methods Section, Statistical Society of Canada (in press).

Fuller, W. A., Dodd, K. W., and J. Wang. 1999. Estimation for the 1997 National Resources Inventory. Unpublished manuscript, Statistical Laboratory, Iowa State University. 161pp.

Nusser, S. M. 1999. Geostatistical Control Data for the 1997 National Resources Inventory. Report submitted to the Natural Conservation Service under cooperative agreement numbers 68-3A75-8-62 and 68-7482-8-351.

U.S. Department of Agriculture. 1997. Instructions for Collecting 1997 National Resources Inventory Data. www.ftw.nrcs.usda.gov/nri/inst_toc.html