Mining GIRAS: Improving on a national treasure of land use data

Curtis Price, Naomi Nakagaki, Kerie Hitt, and Rick Clawges
U.S. Geological Survey

Abstract

Historical land use and land cover data, available from the U.S. Geological Survey (USGS) for the conterminous United States and Hawaii, were converted by USGS and the U.S. Environmental Protection Agency (USEPA) to a geographic information system format in the early 1990's. These data have been available on USEPA's Web site for a number of years and have been used for many national applications, despite minor coding and topological errors. During the 1990's, a group of USGS researchers made edits to improve the data for use in the National Water-Quality Assessment Program. These edited files have been recently modified to create a more accurate, topologically clean, and seamless national data set. Several different procedures, including custom editing software and several batch processes, were required to create this best-available version of this national data set. The next step is to complete the review of this data set to allow the data to be shared with the wider GIS community.

Introduction

During the mid 1970’s and 1980’s, the USGS produced nationally consistent maps of land use and land cover (LULC) for the conterminous U.S. and Hawaii. Polygons of land use and land cover were delineated manually from aerial photography and mapped following a two-level hierarchical classification system (table 1) as described by Anderson and others (1976). The minimum mapping unit used was 4 hectares for all urban and water classes and a few other categories such as "confined feeding operations" and "strip mines, quarries and gravel pits," and 16 hectares for all other classes. The USGS published these land use and land cover maps at USGS 1:250,000- and 1:100,000-scale for the conterminous United States and Hawaii.

The USGS developed the Geographic Information Retrieval and Analysis System (GIRAS) software (Mitchell and others, 1977) to digitize, edit, and produce cartographic and statistical output from the mapped information. The acronym "GIRAS" has been used informally over the years to refer to the software system, its digital file storage format, and the land use and land cover data set (only one of many GIRAS layers) created using the software. The data are currently distributed by the USGS in the original GIRAS file format on the World Wide Web (U.S. Geological Survey, 1986, 1998). For the purposes of this paper, "GIRAS" will be used to refer to the digital GIRAS-format files and “LULC” will refer to the land use and land cover polygon data. A map showing a LULC data set for a single 1:250,000-scale quadrangle is shown in figure 1.

In the early 1990's, the USGS developed a procedure to convert the LULC data set to ArcInfo[1] polygon coverage format so it could be easily used for GIS analysis supporting water-resources applications. USEPA also was interested in using this data in ArcInfo format and worked with USGS to convert all of the quadrangles using the USGS-developed software. The conversion process included: 1) converting the data from GIRAS to ArcInfo coverage format, 2) scaling the data to true Universal Transverse Mercator (UTM) coordinates from the 16-bit coordinate system used in the GIRAS files, 3) projecting to a single national projection, 4) inserting 1:100,000-scale information where available to create a single combined data set for each 1:250,000-scale quadrangle tile, and 5) clipping the data to the quadrangle boundary. USEPA posted the data (in ArcInfo EXPORT format) on its FTP site in 1994, and later, on its Web site. This data set, referred to below as "EPAGIRAS," is still available (U.S. Environmental Protection Agency, 2000).

In the early 1990s, the USGS began to use the LULC data to characterize watersheds and other areas studied by the National Water-Quality Assessment (NAWQA) Program (Gilliom and others, 1995). As NAWQA researchers began working with the LULC data on a national scale, they noticed that minor errors existed in the EPAGIRAS national data set. Some of these minor errors existed in the original GIRAS data files, and some were minor problems introduced in the conversion from GIRAS to ArcInfo format. A comprehensive effort was made at that time to download all the data from USEPA, and fix some of the obvious errors in the data. In the few cases where the ArcInfo files were unsuitable, some of the original GIRAS files were used.

This improved version of the LULC data set has been used in several NAWQA applications. This "NAWQA version" of the LULC data has been used by NAWQA for planning and data analysis. A method was developed to estimate new residential areas by combining the LULC data with U.S. Census population data compiled at the block-group level (Hitt, 1994). Gilliom and Thelin (1997) used this information together with agricultural census data from the U.S. Department of Agriculture to map agricultural land to be used in water-quality assessment. Stewart (1998) developed a land cover data update algorithm based on a combination of the GIRAS data with a Landsat-based land-cover classification.

This NAWQA version of the LULC data set still had polygon coding errors and neatline problems where quadrangle boundaries did not exactly fit together. Because the data are more useful in a corrected, seamless form, the USGS implemented GIS processing steps to create a national seamless data set with better land use code accuracy. The purpose of this paper is to describe these GIS processing steps.

Summary of processing steps

The data processing flow used to develop an enhanced LULC data set is shown in figure 2, and is described in this section. The steps include wholesale replacement of a few of the quadrangle

[1] The use of firm, trade and brand names in this paper is for identification purposes only and does not constitute endorsement by the U.S. Government.

data sets, edits to the polygon LULC codes, data format conversions, and the development of a polygon data set documenting each quadrangle tile of LULC data.

GIRAS data editing

Four quadrangles (Mitchell, S.Dak., Scranton, N.J./Pa., Palestine, Tex., and Cedar City, Utah) either were missing in the posted EPAGIRAS data set or had land cover codes that had been corrupted during the processing flow. The Mitchell, S. Dak., quadrangle data were converted again from the raw GIRAS files. The line work in the Scranton N.J./Pa. data was consistent, but the land cover values in the GIRAS file did not translate into ArcInfo correctly. The paper copy of the USGS published 1:250,000-scale LULC map of the Scranton quadrangle was then used as a backdrop to assign proper land use codes to the polygons. One-quarter of the 1:250,000-scale Cedar City, Utah, quadrangle had no data in the USGS GIRAS file, so the USGS 1:100,000-scale LULC data for that void area were inserted. The four quadrangles were included in the collection of data sets, resulting in complete coverage for the conterminous 48 States and Hawaii.

Coding errors existed in the EPAGIRAS data set that were easily identified by visual inspection. Some of these errors reflected missing labels or polygon boundaries in the original GIRAS file, and some resulted from labels being moved as part of the conversion process from GIRAS to ArcInfo. An example of a polygon that is miscoded (labeled "0") is shown in figure 3.

Arc Macro Language (AML) software was developed to aid in interactive checking and repair of LULC data for obvious coding errors. The older Arc Workstation software (Environmental Systems Research Institute, 2002) was selected for this application because of its ease of use for simple editing and its robust support of coverage editing than newer ArcGIS software. The application allowed users to navigate with a few mouse clicks through all the polygons that had missing or incorrect land use codes and to correct them. The application also included buttons to display two reference data sets in the background: a point coverage of all of the original labels in the GIRAS files (unmodified by GIS polygon processing operations) and the 1992 National Land Cover Data set (Vogelmann and others, 2001). In the rare situation where the proper land use was not clear from the context of the land use of the surrounding polygons, the reference data sets were used to help select an appropriate code. A form menu from the AML application is shown in figure 4. A few of the land use coding errors caused by missing polygon boundaries required the editing of the polygon boundaries to split a single polygon into one or more parts.

A non-interactive process, also in the form of an AML program, was developed to refine the "neatline" processing that had been done earlier to the GIRAS data. This processing basically re-clipped the data to the North American Datum of 1927 (NAD27) quadrangle boundary and extended the polygons to the quadrangle boundary to fill any gaps. The polygons were extended using a hybrid raster/vector algorithm to extend the land use polygons by filling to the quadrangle boundary. The result was a set of GIRAS quadrangle coverages that exactly fit to the intended 1:250,000-scale quadrangle boundaries. This processing was required to allow the quadrangles to fit together with no gaps or overlaps, so the coverages could be easily combined (fig. 5).

Data projection and format conversion

After the LULC data editing was complete, the data were converted and projected into three GIS formats including: 1) ArcInfo coverages in Albers Equal-Area projection, North American Datum of 1983 (NAD83) coordinates, for use in ArcInfo Workstation software, 2) shapefiles, in geographic (latitude-longitude) NAD83 coordinates for use in ArcView 3.x and ArcGIS, and 3) ArcInfo GRID format, in Albers Equal-Area projection, NAD83 coordinates, for use in raster overlay applications.

The individual edited LULC quadrangle coverages were projected to a common projection (Albers, NAD83). These ArcInfo coverages were then used as the basis to convert to other formats. The LULC 1:250,000-scale quadrangle data set names were replaced with spatially-based names; for example: G39083 is the data set with its lower left corner located at 39 degrees south latitude and 83 degrees west longitude. The data were projected into geographic coordinates referenced to the NAD83 datum and converted to ESRI shapefile format. A simple geodatabase was created that allowed the LULC shapefiles to be loaded directly into a geodatabase feature class (fig. 6).

The set of ArcInfo coverages was converted to a 30-meter grid spacing, using a coordinate system designed to line up precisely with the National Land Cover Data Set for potential applications that combine the two data sets. A 30-meter grid cell spacing captures most of the detail in the LULC data detail (fig. 7). The resulting data set is a single ArcInfo integer grid covering the conterminous United States composed of 97,167 rows by 154,133 columns. Because of the run-length encoding in the ArcInfo integer grid data sets, the size of this grid stored on disk totaled 545 MB. The grid had to be stored as an integer grid because the floating point ArcInfo grid format is uncompressed, and the total file size would be very large (approximately 57,130 MB). A file of this size would not only be difficult to manage, but would also exceed the 2.1 MB maximum file size supported by most operating systems.

Metadata information describing each of the original GIRAS files was stored in the GIRAS file header. Much of this information was captured and encoded by using DOCUMENT.AML software (Environmental Systems Research Institute, 2002) as part of the original USEPA/USGS conversion process. These metadata files were processed to extract the metadata information and aggregate it into a single table. In addition, a polygon data set depicting the extent LULC quadrangle data sets was generated. Figure 8 shows an example of how metadata information is associated with each polygon. Three Federal Geographic Data Committee (FGDC) standard metadata files describing the entire data set were created for the coverage, shapefile, and 30-meter grid LULC data sets.

Results

The results of this processing are the following products:

1) A set of LULC coverages in Albers Equal-Area projection, NAD83 datum, with improved land use coding and no overlap or gap between different quadrangle tiles;

2) A set of LULC shapefiles in geographic coordinates, NAD83 datum, easily usable by ArcView 3.x and ArcGIS software;

3) A 30-meter resolution ArcInfo grid in Albers equal-area projection, NAD83 datum, with grid cells aligning with the National Land Cover Data Set for use in regional and national GIS analysis;

4) An ArcGIS geodatabase designed to easily load the shapefile data files and containing a polygon feature class documenting information specific to each quadrangle; and

5) FGDC metadata files describing each of the above data sets.

The next step is to complete the documentation and review of this data set to make this data available to the wider GIS community. Users interested in assisting in the review of this data set are encouraged to contact Curtis Price (cprice@usgs.gov).

Acknowledgments

The authors would like to thank Ed Partington (USEPA) and Doug Nebert (USGS) for their commitment and hard work converting the GIRAS files into ArcInfo format in the early 1990's, and Bernard McNamara (USGS, Sacramento District) for his assistance in aggregating and managing the original NAWQA version of the LULC data sets.

End Notes

The following Web links may be useful for users of USGS LULC data:

USGS Land Cover Products

http://edc.usgs.gov/products/landcover.html

USGS Land Use and Land Cover (LULC) data (GIRAS format)

http://edc.usgs.gov/products/landcover/lulc.html

USGS Land Use and Land Cover (LULC) data (ArcInfo EXPORT format)

http://www.epa.gov/ngispgm3/spdata/EPAGIRAS/

References

Anderson, J.R., Hardy, E.E., Roach, J.T., and Witmer, R.E., 1976. A land use and land cover classification system for use with remote sensor data: U.S. Geological Survey Professional Paper 964, 28 p.

Environmental Systems Research Institute, 2002, ArcInfo Workstation ArcDoc: Redlands, Calif. [on-line documentation]

Gilliom, R.J., Alley, W.M., and Gurtz, M.E., 1995, Design of the National Water-Quality Assessment Program--occurrence and distribution of water-quality conditions: U.S. Geological Survey Circular 1112, 33 p., accessed June 10, 2003 at URL http://pubs.water.usgs.gov/circ1112

Gilliom, R.J., and Thelin, G.P., 1997, Classification and mapping of agricultural land for national water-quality assessment: U.S. Geological Survey Circular 1131, 70 p.

Hitt, K.J., 1994, Refining 1970's land-use data with 1990 population data to indicate new residential development. U.S. Geological Survey Water-Resources Investigations Report 94-4250. 12 p., accessed June 10, 2003 at URL http://water.usgs.gov/nawqa/WRI94-4250.html

Mitchell, W.W., Guptill, S.C., Anderson, K.E., Fegeas, R.G., and Hallam, C.A., 1977, GIRAS—A geographic information analysis system for handling land use and land cover data: U.S. Geological Survey Professional Paper 1059, 16 p.

Stewart, J.S., 1998, Combining satellite data with ancillary data to produce a refined land-use/land-cover map: U.S. Geological Survey Water-Resources Investigations Report 97-4203, 11 p.

U.S. Environmental Protection Agency, 2000, U.S. Geological Survey GIRAS land use and land cover data in ArcInfo format, accessed January 1 2003, at URL: http://www.epa.gov/ngispgm3/spdata/EPAGIRAS/ [on-line geospatial data]

U.S. Geological Survey, 1986, Land use and land cover digital data from 1:250,000- and 1:100,000-scale maps: Data User Guide 4, 25 p.

U.S. Geological Survey, 1998, Land use and land cover digital data from 1:250,000- and 1:100,000-scale maps, accessed January 1, 2003, at URL http://edcwww.cr.usgs.gov/products/landcover/lulc.html [on-line geospatial data]

Vogelmann, J.E., Howard, S.M., Yang, L., Larson, C.R., Wylie, B.K., and Van Driel, N., 2001. Completion of the 1990's National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources: Photogrammetric Engineering and Remote Sensing, v. 67, p. 650-662.

Author Contact information:

Curtis Price, USGS, 1608 Mountain View Rd, Rapid City, SD 57702
Tel: 605-355-4560 x242; E-mail: cprice@usgs.gov

Table 1.— Land use and land cover classification system for use with remote sensor data

[Data from Anderson and others, 1976]

Level I		Level II	
1.	Urban or built-up land	11	Residential
		12	Commercial and services
		13	Industrial
		14	Transportation, communications and utilities
		15	Industrial and commercial complexes
		16	Mixed urban or built-up land
		17	Other urban or built-up land
2.	Agricultural land	21	Cropland and pasture
		22	Orchards, groves, vineyards, nurseries,
		23	Confined feeding operations
		24	Other agricultural land
3.	Range land	31	Herbaceous rangeland
		32	Shrub and brush rangeland
		33	Mixed rangeland
4.	Forest land	41	Deciduous forest land
		42	Evergreen forest land
		43	Mixed forest land
5.	Water	51	Streams and canals
		52	Lakes
		53	Reservoirs
		54	Bays and estuaries
6.	Wetland	61	Forested wetland
		62	Nonforested wetland
7.	Barren land	71	Dry salt flats
		72	Beaches
		73	Sandy areas other than beaches
		74	Bare exposed rock
		75	Strip mines, quarries, and gravel pits
		76	Transitional areas
		77	Mixed barren land
8.	Tundra	81	Shrub and brush tundra
		82	Herbaceous tundra
		83	Bare ground
		84	Wet tundra
		85	Mixed tundra
9.	Perennial snow or ice	91	Perennial snowfields
		92	Glaciers

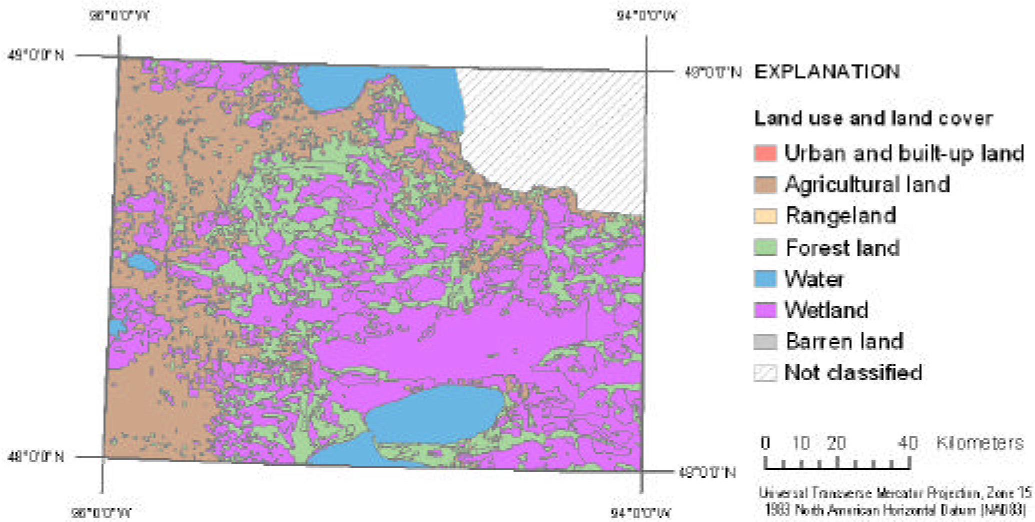

Figure 1. Land use and land cover (generalized) for the Roseau, Minnesota 1:250,000-scale quadrangle

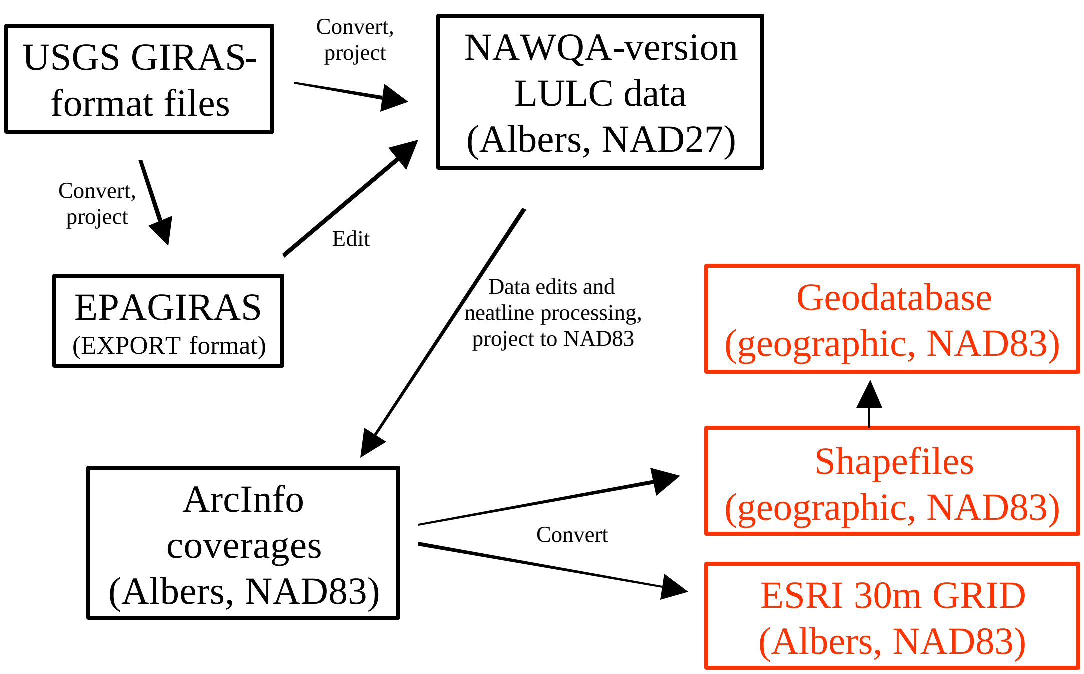

Figure 2. Data processing flow for land use and land cover data

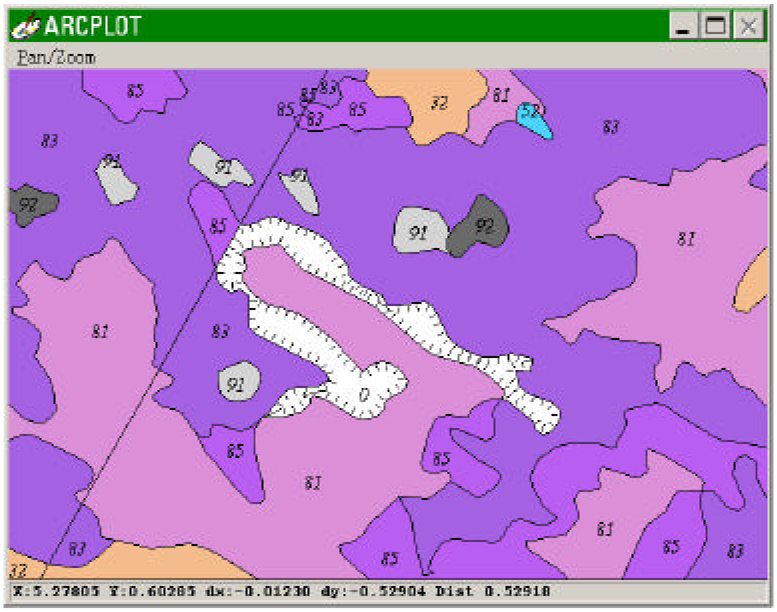

Figure 3. An example coding error in land use and land cover data. The unfilled polygon labeled “0” should be labeled “85.”

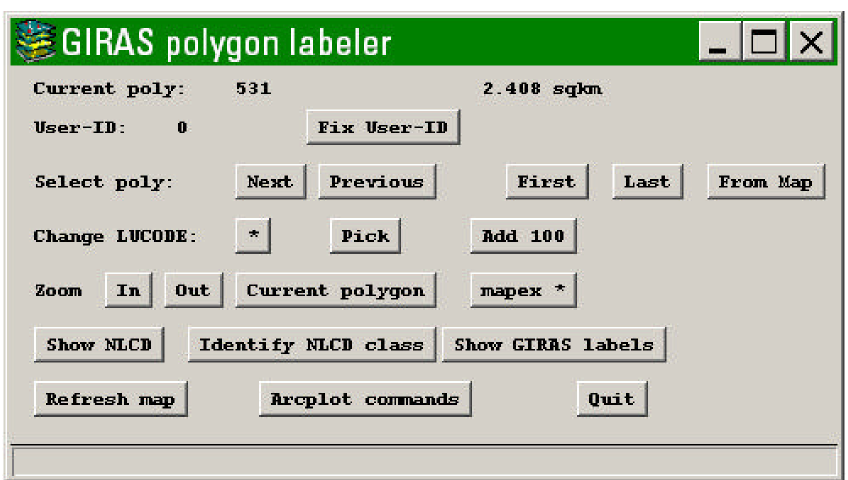

Figure 4. Form menu from land use and land cover editing application

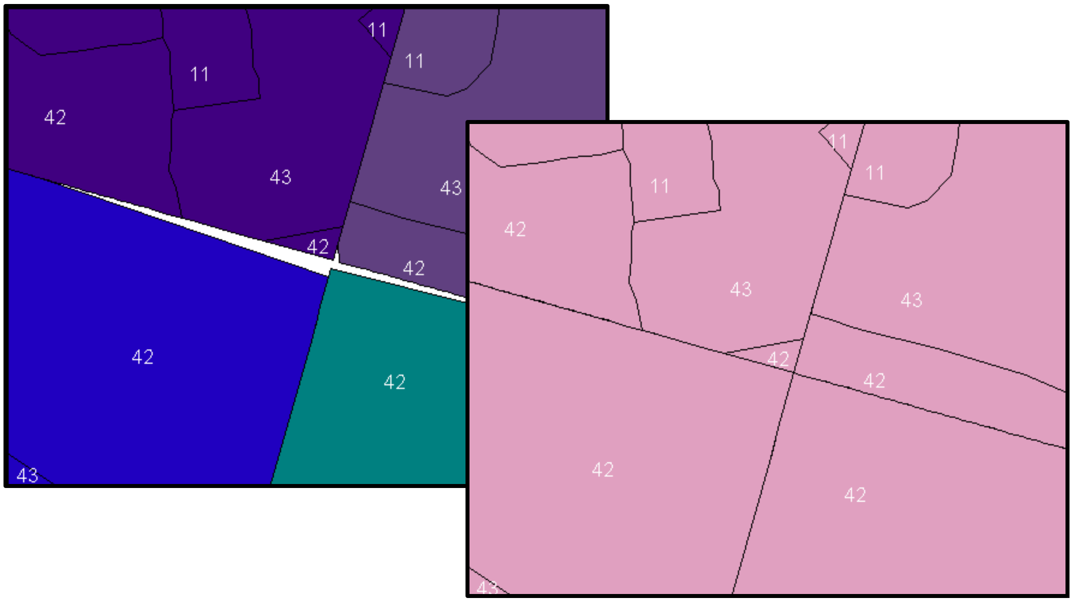

Figure 5. Example of neatline processing of land use and land cover data

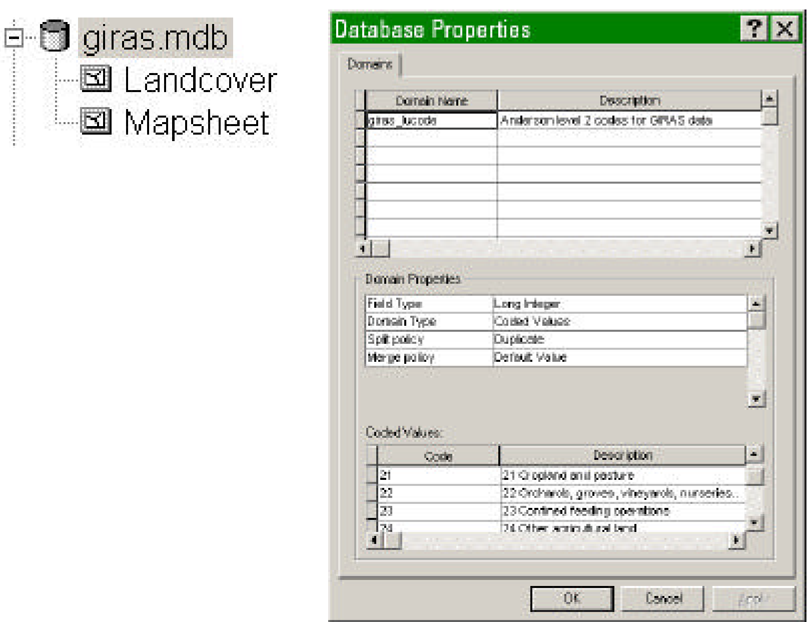

Figure 6. Example showing an ArcCatalog view of a geodatabase model for USGS land use and land cover data

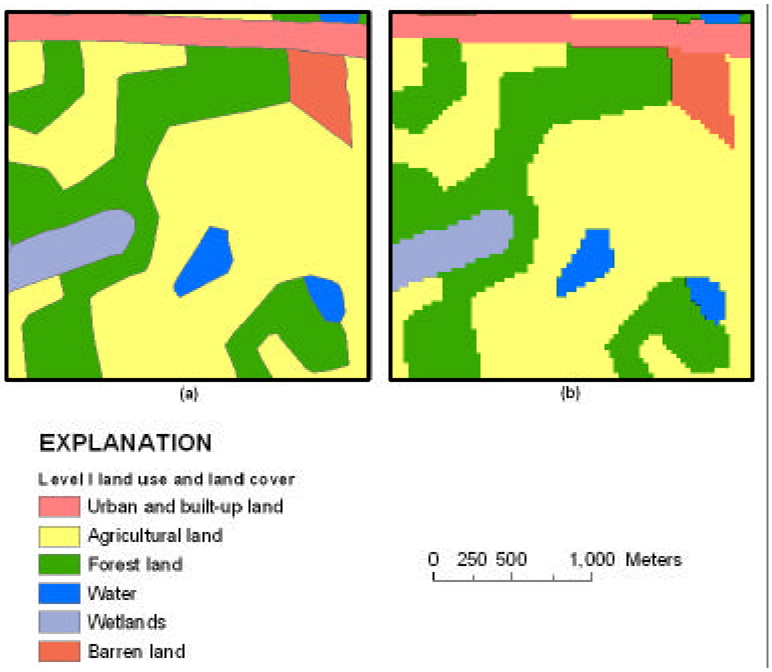

Figure 7. Example of (a) land use and land cover polygons (b) gridded to a 30-meter cell resolution.

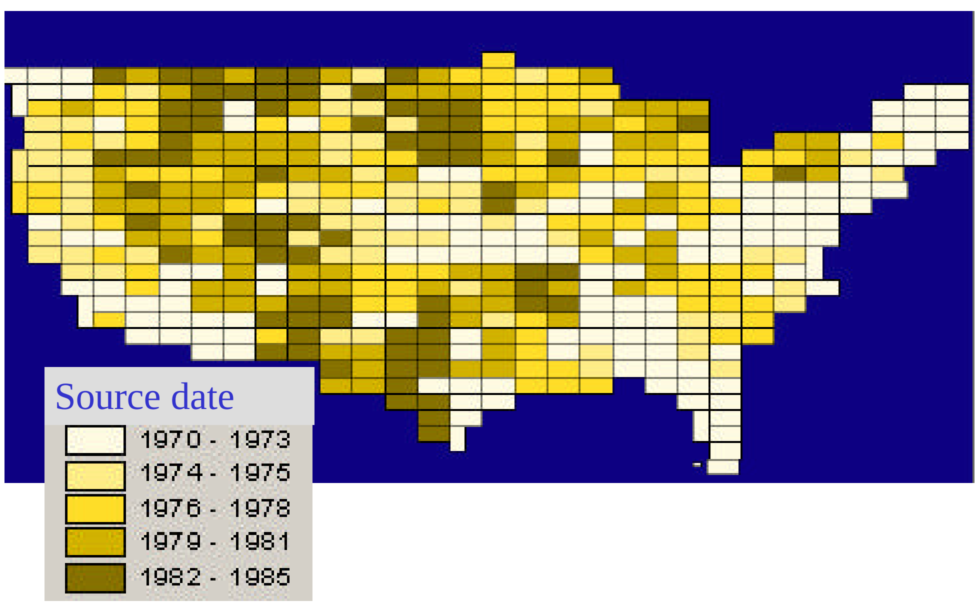

Figure 8. Map showing land use and land cover metadata source dates for quadrangle tiles in the conterminous United States