USGS Experience in Constructing a Global Database

Bradley C. Reed

Abstract

A global land cover database was developed for use in a wide variety of applications, including modeling, sustainable development, and operational tasks. The database was based on 1992-93 1-km resolution Advanced Very High Resolution Radiometer (AVHRR) satellite data supplemented with other earth science information. The classification methodology was based on an unsupervised classification with extensive interactive post-classification refinement. The resulting land cover data have been translated into several conventional land cover schemes that are commonly used in many applications. The land cover data have been validated using a stratified random sample of higher resolution imagery. Results indicate accuracy in the order of 60-75% overall with higher accuracy reported in tropical and subtropical regions and less accuracy in temperate and boreal zones.

Introduction

The U.S. Geological Survey (USGS) EROS Data Center (EDC), in partnership with several international agencies and universities, produced a global land cover characteristics database. The project was part of several programatic initiatives, including the USGS Global Change Research Program, NASA Pathfinder Program, and the International Geosphere Biosphere Programme-Data and Information System (IGBP-DIS). The impetus to develop the 1-km global database grew from the requirements of numerous scientific organizations (*e.g.* U.S. National Academy of Sciences, National Aeronautics and Space Administration) for improved land cover data for global change research, assessments of sustainable development, and operational functions (*e.g.* weather forecasting).

The EDC, in cooperation with numerous sources, led the effort to create the first global 1-km land cover characteristics database. The database was designed to maximize flexibility for use in a number of applications, for ease of use, and to incorporate user feedback to improve the database in subsequent revisions.

Role of cooperators

Cooperators filled a number of roles in contributing to the development and validation of the global land cover characteristics database. Approximately twelve institutions contributed data to support this project and eighteen institutions contributed their expertise in interpreting land cover regions. This included on-site (to EDC) visiting scientists, visiting scientists of institutions supporting EDC, and reviewers of preliminary versions of the database. Each of the visiting scientists was a recognized expert in vegetation or land cover studies over a particular ecological region. In addition, four workshops were held to bring together regional experts to assist in satellite imagery interpretation. During the validation stage of the study, a panel of experts convened at EDC for two weeks to execute the validation plan as developed by the IGBP-DIS. This included experts from all six populated continents. Finally, the user community at-large continues to play a role in improving the database in upcoming revisions. Through the global land cover characterization (GLCC) web site, user suggestions are solicited and recorded for review by the GLCC team. A revision is expected to be completed and available to users by the end of September 1999.

Raytheon ITSS, Inc., USGS EROS Data Center, Sioux Falls, SD, USA 57198

Role of the international geosphere-biosphere programme (IGBP)

The IGBP served as the science oversight for this project. Several IGBP initiatives identified improved land cover data as necessary for improved studies of global climate change. They identified several shortcomings in most extant land cover data, including;

Inconsistent accuracy
Inappropriate class definitions
Emphasis on potential natural vegetation, rather than actual land cover
Variable source data quality
Outdated data

The IGBP-Data and Information System (DIS) recognized the need for improved global land cover information and organized the Land Cover Working Group (LCWG) to address these needs. The Land Cover Working Group was instrumental in organizing the global land 1-km AVHRR project. Under the auspices of the IGBP-DIS and endorsement from the Committee on Earth Observation Satellites (CEOS), the MODIS land science team, and other scientific organizations, the global AVHRR 1-km database was developed. With the cooperation of over thirty ground-receiving stations, the USGS EROS Data Center assembled a global 1-km NDVI database from 1992-1993 that was used in the construction of the land cover characteristics database. Since then, the global 1-km AVHRR project has continued to produce data free of charge for users.

Through user requirement forums, the LCWG developed a set of specifications for the proposed global land cover database. These specifications are as follows:

1) The database should be of moderate spatial (1-km), and coherent temporal resolution (corresponding to a specific baseline period, such as a 1-year window).
2) The methodology should be objective, reproducible, and systematic.
3) Results should be sufficiently flexible to permit use in a wide range of applications.
4) The land cover legend should be compatible with past, present and future legends to permit comparisons.
5) The data should be comprised of information on seasonal and interannual vegetation dynamics.
6) The database should include socioeconomic, cultural and natural factors that affect global land cover patterns.
7) The project should include validation using a statistically sound accuracy assessment protocol.

Database development techniques

The approach to developing the land cover characteristics database was adapted to conform to the requirements set forth by the LCWG. The database is derived from 1-km AVHRR data spanning the 12-month period from April 1992-March 1993. The data are based on a flexible database structure and a seasonal land cover region concept. Seasonal land cover regions provide a framework for presenting the temporal and spatial patterns of vegetation in the database. The regions are composed of relatively homogeneous land cover associations (for example, similar floristic and physiognomic characteristics) that exhibit distinctive phenology (onset, peak, and duration of greenness) and have common levels of primary production.

Ten-day normalized difference vegetation index (NDVI) data were the core data set used for the land cover characterization. The ten-day composites were re-composited to monthly composites to reduce data volume and to further reduce atmospheric noise. The monthly composites were inserted into an unsupervised classification, continent-by-continent as the initial stage in the classification process. The concept is that by using an unsupervised classification, pixels with similar phenology would be grouped together. Each of the continents was processed separately to reduce confusion between clusters with similar phenology, though entirely different environmental conditions, being grouped together.

The resulting clusters were then interpreted, and assigned a preliminary label, on the basis of their seasonality and their coincidence with any available ancillary earth science data. The ancillary data that were analyzed concurrently with the clusters included global digital elevation model (DEM) data, ecoregions, and any and all available digital, map, or tabular data related to land cover and vegetation. The collection of ancillary data was uneven throughout the globe and was largely based on library holdings in the U.S. and in cooperators' home institutions and libraries.

The preliminary greenness classes were processed into seasonal land cover regions (SLCRs) using post-classification refinement involving iterative splitting and joining of clusters. The interpretation was based on extensive use of computer-assisted image processing tools, but the process was not entirely automated and resembled traditional manual image interpretation. Interpreter skill was required to make final decisions regarding the relationship between the temporal spectral classes and landscape characteristics exhibited by the ancillary data to produce the final land cover definitions. Using all available documentation, a convergence-of-evidence approach was used by a team of interpreters to assign the SLCR label.

The SLCR label does not conform to any pre-defined land cover legend. Rather it is a free-form description of what the interpreters consider to constitute a particular cluster or sub-cluster. A typical label might be "mixed deciduous forest with pastures and cropland" or "douglas-fir, hemlock, oak" or "bluestem, wheatgrass, small grains, row crops". While these labels may not fit into a particular land cover legend, they do provide the basis for translation into a number of legends.

The mechanism for legend translation was based on the Global Ecosystems (GE) legend. The GE legend is composed of 94 ecosystem classes that are based on land cover mosaics, floristic properties, climate and physiognomy. The SLCRs were translated into their appropriate GE class which allowed for tailoring data to the unique landscape conditions of each continent, while still providing a means for summarizing the data at the global level. The ecosystem types were then cross-referenced to land cover classes in the legends as defined by the Simple Biosphere Model, the Biosphere Atmosphere Transfer Scheme, USGS/Anderson, and the IGBP.

Delivery of products

The continental data layers were joined to form global land cover data sets. The global data sets are presented in a generic binary format in the interrupted Goode Homolosine Equal Area projection, and the continental data sets are available in both the Goode and Lambert Azimuthal Equal Area projections. The GLCC project team will make the global data set available in geographic coordinates with the next revision. All the data used or generated during the course of the GLCC development, unless protected by copyrights or trade secret agreements, are part of the final database that is available free of charge via the Internet at: *http://edcwww.cr.usgs.gov/landdaac/glcc/glcc.html.*

User profile

The GLCC web site includes a voluntary User Registry. The User Registration includes information on the users' origin, their applications, and their study area. This was meant to provide a description of database users so that their needs could be better met and to keep users informed about updates and other announcements. Slightly under 50% of the over 600 registered users are from the United States and nearly 25% are from Europe. The database has been accessed from 57 countries.

Applications using the land cover database vary widely. They range from applications for which the data were originally designed, such as modeling global climate change scenarios, to uses that were not anticipated, such as analyzing migratory bird flyways. Applications can be grouped into four main categories: mapping, modeling, environmental management, and general interest. The largest group of users is involved in modeling

(~50%) and environmental management. Approximately 20 % of the users have an interest in mapping applications and 15 % in environmental management, and 15 % have a non-scientific interest in the data.

Because of the paucity of land cover information for much of the Earth, the availability of the database has sometimes become a means for filling a void that it was not meant to address. The data have been used for applications on regional and local scales that they were not meant to serve. But the overwhelming majority of users are, indeed, using the data for continental and global applications.

Validation

The IGBP-DIS Land Cover Working Group also formed a sub-group on validation. The Validation Subgroup had the task of balancing what was desired with what was practical. Since the objective was to provide accuracy figures on the global land cover product bearing the IGBP legend (DISCover), the sampling strategy was based upon that product.

The sampling strategy was a stratified-random approach with 25 random samples (pixels) collected for each of 15 of the 17 DISCover classes (water and snow/ice classes were not validated). Three independent interpreters determined the "true" cover types by analyzing high-resolution satellite data covering each sample.

Accuracy figures were reported in two ways. One method involved the majority rule, where there was a consensus between the 3 interpreters, and resulted in a 74.3 % overall accuracy. In another approach, if the 3 interpreters all disagreed, the data were reported as inaccurate. This happened in 69 of the 375 cases and resulted in a 59.4 % accuracy rating. Another approach was to use an area-weighted accuracy where the figures were adjusted according to the areal coverage of the land cover type. The accuracy was 78.7 % and 70.5 % for the first two approaches, respectively.

Lessons learned

Data quality affects results. Atmospheric contamination, despite compositing on a monthly time scale, caused significant problems, especially in heterogeneous landscapes in the subtropical, temperate, and boreal biomes. In addition, the availability of reference data limited the ability of interpreters to label clusters in many remote areas. Moreover, the quality and timeliness of some of the ancillary data were not optimal.

With more time, a better product could be created. The IGBP-DIS required delivery of the data to the LCWG validation team by 1 July, 1997. The first 12-month set of AVHRR composites for North America was completed in March 1995, at which time the mapping process began. Due to the availability of AVHRR data, the period for mapping North and South America was longer than that for Eurasia and the Australia-Pacific area. In addition, due to the short mapping period, those continents also had a short time for external review before the final deadline.

Resources have a significant effect on results. Of all the factors involved in constructing the database, staff and budget resources may have the greatest impact on the quality of results. The project required approximately 10.5 staff-years, spanning 27 months. In addition, nearly 2.0 staff-years were devoted to gathering reference materials, ensuring the quality of AVHRR composites, and other preparatory tasks.

User interaction is essential. At many different levels, users' interaction is necessary for developing a large-area database. This includes user input and feedback into the database development as described earlier. It is also essential to provide user assistance in applications involving the database. Frequently, sophisticated scientists are using this type of data for the first time and have difficulty in adapting the data to their specific projects. By working hand-in-hand with many users, we have found that the database was much more adaptable than we had originally hoped.