112291
draft for 031193 ?

Design and implementation of a soil geographic database for rural planning and management[1]

Authors: *R. Norberto Fernández[2], Marek Rusinkiewicz[3], Lucia Morais da Silva[4], and Chris J. Johannsen[5]*

[2]Resource Information Manager, Laboratory for Applications of Remote Sensing (LARS), Purdue University, 214 ENTM Building, West Lafayette, Indiana 47907.

[3]Associate Professor, Department of Computer Science, University of Houston, 4800 Calhoun Road, Houston, Texas 77004.

[4]Graduate Student, Department of Agronomy, Purdue University, Lilly Hall of Life Sciences, West Lafayette, Indiana 47907.

[5]Director, Laboratory for Applications of Remote Sensing (LARS), Purdue University, 214 ENTM Building, West Lafayette, Indiana 47907.

[1]Contribution from the Agronomy Dep., Purdue Univ. Agric. Exp. Stn., West Lafayette, IN 47907. Journal paper no. 12903.

ABSTRACT: The physical design of a database involves the evaluation of implementation alternatives using the data model of the database management system (DBMS). The results of the conceptual design of a soil database were mapped to the relational data model. The resulting database is free from update anomalies (i.e. each elementary fact can be updated independently of other elementary facts), while preserving all dependencies among attributes. The database was implemented using a microcomputer-based DBMS, and loaded with data provided by the Soil Conservation Service (Forms 5 and 6). This database, which is part of a geographic information system (GIS), will provide information for soil erosion and soil management studies, and land appraisal for tax assessment, at county level. To facilitate data retrieval, pre-defined queries have been developed to retrieve data based on various combinations of attributes. The results of queries can be presented as formatted reports, and linked to the cartographic database of the GIS.

Introduction

Geographic information systems (GIS) provide scientists, professionals, managers and decision-makers with an efficient way of combining and analyzing georeferenced and descriptive data from different sources (soils, vegetation, geology, land cover, and others), for a better understanding and management of our natural resources *(22)*.

Geographically referenced, along with descriptive soil data are required for a number of applications in different fields and at different levels of detail *(18)*. Soil data gathered by the Soil Conservation Service at the detailed level, from 1:15,840 to 1:31,680, provide adequate information for rural land planning and management at the county level *(21,22)*. The availability of these data in the US make soil surveys a necessary layer of information in land planning.

As the amount of data to be included in an attribute database and the number of users of the database increase, the need for a structured approach to database design arises. The goals of the database design process are to ensure an efficient data processing through the elimination of redundant information, and the minimization of update and deletion problems *(12)*.

The objective of this study was to design and implement a soil database to be used in rural planning and management, within a microcomputer-based GIS/land information system (GIS/LIS). This initiative is part of an on-going pilot project in Miami County, Indiana, conducted by the Laboratory for Applications of Remote Sensing (LARS), Purdue University, with the purpose of developing a GIS/LIS to solve rural needs. This project is being carried out in cooperation with several local, state and federal agencies, as well as Purdue researchers and graduate students *(13)*.

Components of a geographic information system

A GIS is a computer-based system that is used to store and analyze geo-referenced information *(1)*. Basically, a GIS can be viewed as the organic integration of several subsystems *(2)*: *a) the input subsystem,* which allows the conversion of data to a format that the computer can understand; the input of spatial data, such as maps, is done through digitization; *b) the data management subsystem,* which performs all the data handling operations, and provides the link between data stored in the databases and the models used for analysis; *c) the database subsystem,* where the spatial and attribute data are stored: the spatial data are geo-referenced to a common base map; and the attribute database contains descriptive data related to each layer of the spatial database; *d) the analysis and modeling subsystem* deals with all the analytical operations (boolean algebra) needed to meet the requirements of the user(s); and *e) the output subsystem,* which generates different products that result from analysis, such as maps, reports or statistics (Figure 1). In a more general perspective, a GIS

On the other hand, models used in the physical design stage show how the data will be stored in the database. These models are database management system (DBMS) dependent; thus, most of the abstraction is lost in this part of the process because the data models describe, in this case, implementation details that depend on the DBMS.

General design methodologies developed for non-spatial databases were later integrated with geographic information in formulating a new composite methodology for designing spatial databases *(3)*.

We have chosen the EER model to design conceptually a soil database that is part of a PC-based GIS/LIS for rural planning and management. This database stores selected soil data acquired by the Soil Conservation Service (SCS) for Miami County, Indiana, at the detailed level, Forms 5 and 6 *(7)*.

Several entities and relationships were defined to represent the data and relations among data that are known to this database; these concepts are graphically expressed by the EER schema (Figure 2). After a soil map is digitized and entered into the GIS, the *soil delineations* become polygons; this was indicated by the **MAP-POLYGON** and **TERRAIN** entities, i.e. a polygon of a digitized soil map represents certain terrain characteristics. Land areas classified as *soil consociations* or *soil complexes* in the Miami County soil suvey *(7)*, were represented by the entity **CONSOCIATION/COMPLEX**. Data concerning soil profiles and soil horizons were grouped around the entities **SOIL-INDIVIDUAL** and **LAYERS**, respectively. A similar procedure was followed to represent inclusions of soils, constituents of a soil complex, and the representative soil used to explain a soil consociation (see the relationship types **CONSISTS_OF**, **COMPOSED_OF**, and **HAS_INCLUSIONS_OF** in Figure 2).

All definitions of entities, relationships and descriptive attributes were done within the constraints of the EER model. A detailed description of the conceptual design of this database has been given by Fernández and Rusinkiewicz *(11)*.

The physical design. The database has been implemented, using the relational model *(6)* and the PC INFO[1] software. This attribute database is part of a GIS/LIS developed with PC ARC/INFO[2]. Informally, a relational database is perceived by the user as a collection of tables and nothing but tables *(8)*. All data are in these tables, and all database operations work from these tables.

We have derived a relational schema from the conceptual schema shown in Figure 2, according to existing rules for conceptual-relational mapping *(9)*. A mapping between the conceptual design and the physical design converts all

---

[1] INFO is a trademark of Henco Software, Inc., 100 Fifth Avenue, Walthman, MA 02154.

[2] ARC/INFO is a trademark of Environmental Systems Research Institute, Inc. (ESRI), Redlands, California.

entities and relationships with attributes to relation schemas. The result is a collection of relational schemas showing all possible relations (tables) that will be implemented using the DBMS.

The following relations were derived from Figure 2: ***MAP-POLYGON, TERRAIN, CONSOCIATION/COMPLEX, CONSISTS_OF, COMPOSED_OF, SOIL-INDIVIDUAL*** and ***LAYERS*** (Figure 3). Even though the data were split in different tables, relations still exist and tables are connected by key attributes. These linking keys for the entire database are **Polygon-ID** and **Soil-ID**. The **Polygon-ID** provides the link between each polygon of the digitized soil map and the descriptive or attribute data associated with it. The second part of this link is supplied by the **Soil-ID**, which is an unique code number that was given to each soil mapped for Miami county. Data associated to each soil can be accessed through the **Soil-ID**.

Because of constraints of the relational model and design options, some of the tables were designed to be accessed through a combination of two key attributes rather than one. This is the case for the relations ***CONSISTS_OF, COMPOSED_OF,*** and ***LAYERS***. For example, the table ***LAYERS*** contains information on physical and chemical characteristics of each soil horizon. To access this table it is necessary to identify first the soil, and then the horizon about which the information is required. This is done through the combination of the key attributes **Soil-ID** and **Layer-ID**. ***LAYERS*** is linked to the rest of the database through the attribute **Pedon-ID** (Figure 3).

One of the objectives of this design was to maximize the number of relations with only one key attribute; this guarantees tables with few attributes and occurrences (tuples). The advantages of this approach are: less storage space required, fast access to tables when many joins are not needed, and flexibility for expansions and future implementations. This results in an improvement of the overall efficiency of the system.

We chose to store data under the ***CONSOCIATION/COMPLEX*** and ***LAYERS*** relations in three and two different tables, respectively. This fragmentation was done according to the users' data processing needs, defined in the early stages of the design. The primary key attributes of each relation are the same, thus, each full relation can be easily reconstructed from the fragments (Figure 3).

The relations ***CONSISTS_OF*** and ***COMPOSED_OF*** are the connections between the lower part of the database, which contains data on soil profiles and horizons, and the upper part which holds data on surface horizons, general soil characteristics and map polygons.

The relation ***MAP-POLYGON*** contains data about each polygon on the soil map. This relation is generated by ARC after a map is digitized and topology is built. The exception is the attribute Soil-ID, which was added to link this relation to the rest of the database.

The diagram defined during the physical design (Figure 3) displays only some aspects of the physical schema; thus, some refinements are needed for the final storage structure. For example, soil data (in SCS databases) are coded. This codification is a normal procedure in database implementation in order to minimize storage space, and to provide a better organization of the data. But code numbers, in general, have no meaning to the user(s); therefore, additional tables (look-up tables) are created to "translate" these numbers into words and phrases that the user can understand.

In this scheme, soil data have been stored in the database using the same coding system as SCS. We have also implemented look-up tables for data display purposes. For example, when information about the erosion class or the organic matter (OM) content of the surface horizon of the "Fox Clay Loam, 8-15% slopes" soil is requested (Figure 4), the DBMS will read the code numbers for the attributes, and then will translate them into words through the look-up tables. In this example, the number *1* for the attribute EROSION, will read as "severely eroded"; and the number *3* for OM, will read "moderate".

It is important to notice that the data provided by SCS were preserved, and no changes were made when accommodating them in the database; except for those changes that were imposed by the DBMS we used. Some attributes, though, were disregarded because they were not relevant to the applications of this database, i.e. soil mechanical properties. Descriptions for each code number, in the look-up tables, were adapted from those of USDA/SCS *(20)* for the purposes of this database.

Results

An efficient system must assure a good quality of information and provide a simple way to retrieve that information. To accomplish this, we developed an applications-oriented database, i.e. various programs were written for queries, data update and deletion. For database queries, these macros offer sequential questions that allow the user to select the variable(s) of interest and, then, a range of values for that(those) variable(s). Results can be produced "on-screen", and on "hard copies" in map and report formats. These interactive programs allow a series of operations that range from simple queries with a few variables, up to the combination of soil data with other layers of data for analysis and modeling. The integration of spatial and attribute databases within a GIS/LIS environment provides efficient data handling for rural planning and management (Figure 5 and 6).

Conclusions

We have designed and implemented a soil database as part of a microcomputer-based GIS/LIS to provide information in solving rural needs in Miami County, Indiana. Two data models were used to provide better system efficiency and flexibility: both the EER model for conceptual design and the relational model for implementation of the database, presented excellent results.

After all data are loaded, this database will contain a minimum of 16,000 records in the **MAP-POLYGON** relation, 200 records in the **LAYERS** relation, and 55 records in each remaining relation. Using a database design as the one proposed here, soil data can be efficiently stored and processed with microcomputers without compromising the access time and performance.

This database was designed in such a way that expansions and future implementations can be accommodated without substantial modifications of the basic design. Emerging technologies in database-oriented hardware will also come to establish good organization and fast data access when needed.

Acknowledgements

This research was funded by NASA Grant NAGW-1472. We thank the Soil Conservation Service, Indianapolis Office, Indiana, for providing the digital soil data used in this work. Ms. Morais da Silva acknowledges financial support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) of Brazil.

REFERENCES CITED

1. Aronoff, S. 1989. *Geographic information systems: a management perspective.* WDL Publications, Ontario, Canada. 294 pp.

2. Bartolucci, L. A., T.L. Phillips, and C.R. Valenzuela. 1983. *Bolivian digital geographic information system.* In Proc., Ninth International Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana. pp 374-388.

3. Calkins, H.W. 1982. *A pragmatic approach to geographic information systems design.* In Proc., US/Australia Workshop on Design and Implementation of Computer-based Geographic Information Systems, Peuquet and O'Callaghan [eds.], Amherst, NY: IGU Commission on Geographical Data Sensing and Processing.

4. Chambers, D. 1989. *Overview of GIS database design.* ARC News, Spring, 11(2).

5. Chen, P.P. 1977. *The Entity-Relationship model: toward a unified view of the data.* ACM Transactions on Database Systems, 1(9): 9-36.

6. Codd, E.F. 1970. *A relational model of data for large shared data banks.* CACM, (13)6.

7. Deal, J.M. 1979. *Soil survey of Miami county, Indiana.* U.S. Department of Agriculture, Soil Conservation Service, in cooperation with Purdue University Agr. Exp. Station. 138 pp. and 30 map sheets.

8. Date, C.J. 1988. *An introduction to database systems.* Addison-Wesley, Reading, MA. 639 pp.

9. ElMasri, R., and S.B. Navathe. 1989. *Fundamentals of database systems.* Benjamin/Cummings Publ. Co., Inc. Redwood City, CA. 802 pp.

10. ElMasri, R., A. Hevner, and J. Weeldreyer. 1985. *The category concept; an extension to the Entity-Relationship model.* Data and Knowledge Engineering. 1(1): 75-116.

11. Fernández, R.N., and M. Rusinkiewicz. 1991. *Conceptual design of a soil database for a geographic information system.* Submitted for publication.

12. Jackson, G.A. 1988. *Relational database design with microcomputer applications,* Prentice Hall, Englewood Cliffs, New Jersey. 207 pp.

13. Johannsen, C.J., R.N. Fernández, D.F. Lozano-García, and J. Hart. 1990. *Applying remote sensing and GIS techniques in solving rural county information needs.* Proceedings of the 2nd Annual Indiana GIS Conference, Indianapolis, Indiana, pp. 143-150.

14. Marble, D.F. 1988. *Approaches to the efficient design of spatial databases at a global scale.* In Building Databases for Global Science, Mounsey, H. and R.F. Tomlinson [eds.], Taylor and Francis, London, pp 49-65.

15. Navathe, S., R. Elmasri, and J. Larson. 1986. *Integrating users views in database design.* IEEE Computer, 19(1): 50-62.

16. Navathe, S., and M. Schkolnick. 1978. *View Representation in logical database design.* In Proc., ACM-SIGMOD International Conference on Management of Data, Austin, TX. pp. 144-156.

17. Nielsen, G.A., J.M. Caprio, P.A. McDaniel, R.D. Snyder, and C. Montagne. 1990. *MAPS: a GIS for land management in Montana*. J. Soil and Water Cons. 45(4):450-453.

18. Reybold, W.U., and G.W. TeSelle. 1989. *Soil geographic data bases*. J. of Soil and Water Cons. 43(3):226-229.

19. Soil Survey Staff. 1919. *Map units*. In Soil Survey Manual rev., Chapter 5, Soil Conservation Service-United States Department of Agriculture, Washington D.C., Govt. Printing Office. In press.

20. USDA/SCS. 1979. *PEDON coding system for the National Cooperative Soil Survey*. United States Department of Agriculture/Soil Conservation Service.

21. Ventura, S.J., B.J. Niemann, Jr., and D.D. Moyer. 1988. *A multipurpose land information system for rural planning*. J. of Soil and Water Cons. 43(3):226-229.

22. Ventura, S.J., N.R. Chrisman, K. Connors, R.F. Gurda, and R.W. Martin. 1988. *A land information system for soil erosion control planning*. J. of Soil and Water Cons. 43(3):226-229.

23. Walsh, S.J. 1985. *Geographic information systems for natural resource management*. J. Soil and Water Cons. 40(2):202-205.

List of Figures

Figure 1. Basic components of a geographic information system (adapted from Bartolucci et al., *2*).

Figure 2. Conceptual schema for the soil database. Entities, relationships and attributes are represented with rectangles, diamonds and ovals, respectively. Double-ovals indicate composed attributes (from Fernández and Rusinkiewicz, *11*).

Figure 3. A relational schema corresponding to the conceptual schema of Figure 2 (adapted from Fernández and Rusinkiewicz, *11*).

Figure 4. Example of implementation of look-up tables for coded variables.

Figure 5. GIS query for three variables: available water holding capacity (high and very high), slopes (less than 5%) and organic matter content of the surface horizon (high and very high).

Figure 6. Soil map units that correspond to the GIS query shown in Figure 5, in report format.