Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2000 Proceedings

Americas Conference on Information Systems (AMCIS)

2000

A Dual Census Geographical Information System: Is It a Data Warehouse?

James B. Pick
University of Redlands, pick@uor.edu

Nanda Viswanathan
University of Redlands, viswanat@uor.edu

W. James Hettrick
City of Loma Linda, hettrick@ci.loma-linda.ca.us

A Dual Census Geographical Information System: Is It A Data Warehouse?

James B. Pick, Department of Management and Business, University of Redlands, pick@uor.edu,
Nanda Viswanathan, Department of Management and Business, University of Redlands, viswanat@uor.edu
W. James Hettrick, City of Loma Linda, hettrick@ci.loma-linda.ca.us

Abstract

Geographical Information Systems (GIS) for governmental applications are beginning to be applied on a dual-census or multi-census basis. A dual census GIS is one that references the data of two censuses in an integrated way. Nearly every nation worldwide has a census that constitutes its primary repository of spatially referenced social and economic data. However, censuses vary a lot in their data quality, attribute definitions, and spatial design. This paper has the primary goal to examine to what extent a dual census GIS corresponds to a data warehouse. The paper first presents a contemporary design for a dual census GIS, followed by an example. Then it reviews the key features of a data warehouse and examines feature by feature whether or not a dual census GIS corresponds to a data warehouse.

The paper concludes that a dual census GIS mostly corresponds to a data warehouse. It succeeds on features of combining data from multiple sources, but falls short on multi-dimensionality. A dual census GIS emphasizes calculation and modeling to a greater extent than most data warehouses. In this respect, the dual census GIS meets the OLAP criteria, and in particular that of data manipulation. The paper points out that the real advantage of incorporating data warehousing may be to combine it into a spatial decision support system, in order to provide decision support in a bi-national or multi-national context.

Introduction and Goals

Geographical information systems applications originated in government (Huxhold, 1991) and have been utilized extensively by all levels of governments -- local, state, and national (Longley et al., 1999). One of the major sources of data, if not the most important one, is the national census. Nearly all of the approximately 200 nations conduct censuses, mostly commonly every ten years. Census data are utilized by all levels of government, as well as extensively by private industry. For instance, census data form a important part of many marketing data-bases for small areas.

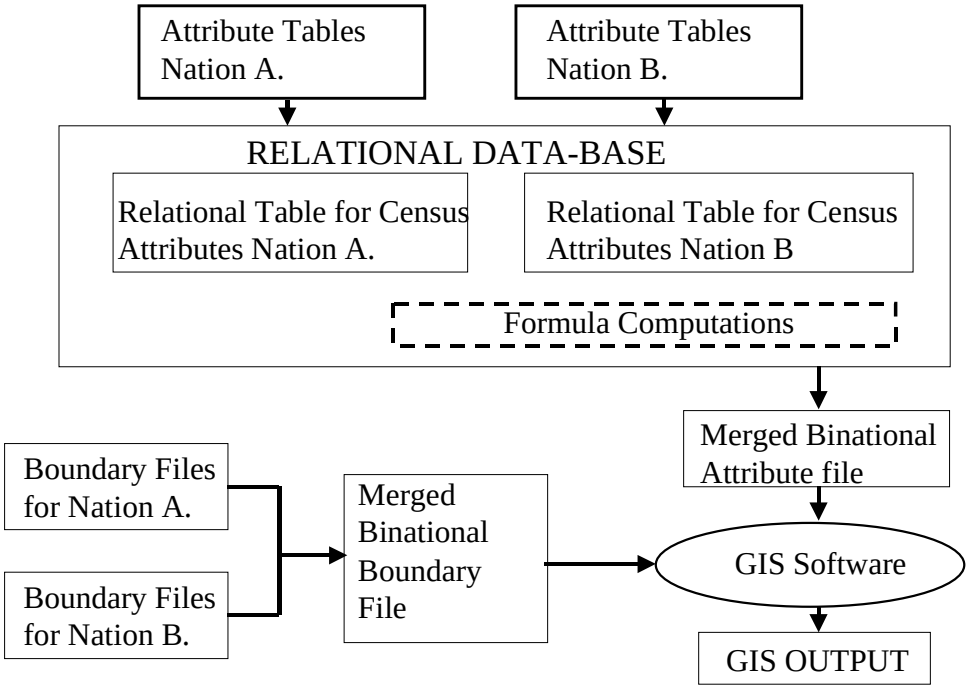
A dual census GIS is one that combines data from two censuses to provide integrated and coordinated access and use (Pick et al., 2000). A multi census GIS combines three or more censuses in the same way. For consistency, this paper refers to dual census GIS, but the results of the paper apply equally well to multi census GISs. An example of a dual census GIS is a GIS for the U.S. and Canada, that combines data from both censuses and enables mapping and analysis of the combined territory of both nations, or of parts of the territory of both nations. The dual census GIS needs to have consistent and equivalently defined geographical units in both nations, such as census tracts or counties, and also to have consistent attribute information, such as income, education, or migration.

An example of a multi census GIS is an integrated GIS for all of the European Union. It includes data from dozens of different censuses that would need to be adjusted into a standard format and arrangement. Another example is based on the Global Urban Observatory Program (UN, 1997). This program is attempting to formulated a standard set of urban indicators for cities globally. A GIS is being planned for the GUO that represents a multi census GIS for the world's major cities, since the urban indicators chosen are largely census-like.

Dual Census GIS -- Introduction

The input components of a dual census GIS, shown in Figure 1, consist of attribute tables and boundary files from censuses of the two nations. In this design, the attribute tables for each nation are extracted from national census data products. The data are adjusted, i.e. cleaned, if necessary, and entered into a relational data-base. Each census has a relational table. Certain attributes will "match" on definition. For instance, total population and gender ratio are two attributes from the two nations that match. However, in many other cases, attribute definitions do not match. For instance income levels do not match, since one might compute income as a category and another nation as a numerical value Attributes that match definitionally can undergo formula computations and be output into the "merged binational attribute file."

Figure 1. Dual-Census GIS Model.



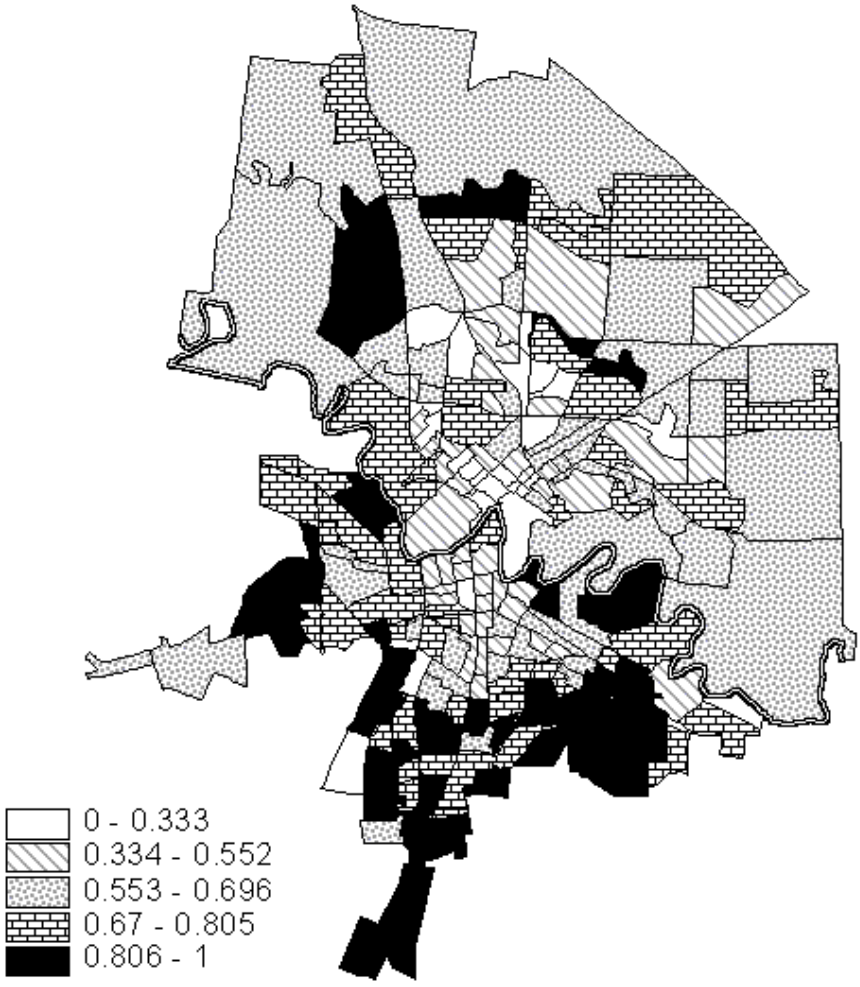
Boundary files present a different challenge for mapping and analysis. For instance, one census's boundary files have Lambert Conformal Conic projection, while another census maintains boundary files in the Albers Equal Area projection. The boundary files need to be adjusted to a common projection. There are many other types of incompatibilities that may exist in the boundary files for the two censuses and may need to be adjusted. For instance, there is an incompatibility in the international boundary line itself i.e. that angles or subtle changes near rivers or cities are slightly different for the two nations. After such incompatibilities are adjusted and cleaned up, the output from these boundary adjustments is a merged binational boundary file.

After the merged binational attribute file and merged binational boundary file have been created, GIS software can be applied to map or analyze the spatial patterns and trends of the two nations in an integrated way.

Dual Census GIS -- Example

The twin cities of the United States and Mexico were analyzed spatially based on a dual census design. The eight twin cities had a year 2000 estimated total population of 6.8 million. A GIS was constructed consisting of 2,958 small areas having an average population of 1739 persons (Pick et al., 1999, 2000). In the U.S., the small area unit utilized was the block group. In Mexico, the small area geographical unit of AGEB (Area Geografica Basica) was used. The average population in the two units is roughly similar. The geographical design is that the U.S. block groups with equal population, whereas the AGEBs are keyed to equally sized areas. However, in the urban context, both can be utilized for mapping without too much spatial incompatibility (see Figure 2).

Figure 2. Example of Display from Dual Census GIS: Percent of Home Ownership for the U.S.-Mexico Twin City of Brownsville, Texas, and Matamoros, Mexico



The attribute tables consisted of 16 "matched" variables. These variables are assigned to four groups, namely population, social, economic, and housing. Some attributes, such as population, dependency ratio, and nativity, were matched exactly in definition, while others including income, and public sewer were matched fairly closely, but not exactly. One attribute, poverty, was not matched closely in definition but was considered impor-

tant enough to include even with weak definitional matching (Pick et al., 1999). Large numbers of attributes from both censuses had to be excluded because it was impossible to match their definitions.

Formula computations were performed using the compute features of MS Access. It was possible to accomplish documentation, by copying the Access formula macro commands into the documentation files. Those files contained additional information as well, including references to the original census data products, attribute definitions, and geographical referencing.

The dual census model in Figure 1 was implemented successfully in the ArcView GIS software for the U.S.-Mexico border cities. It has enabled sophisticated statistical analysis including cluster analysis and estimation of a cluster-based binational index. The software used was the statistical package SPSS and Excel (Pick et al., 1999).

In the future, an enhancement to the dual census GS is to apply longitudinal analysis. For instance, when the year 2000 census results have been made available, it will be possible to compare 1990 and 2000 censuses on a variety of attributes. For instance, educational changes could be recorded for 1990 and 2000. The GIS used for decision support can examine percentage gains in characteristics e.g. poverty, divorce rate, certain building codes. by location at home or office Modeling methods such as spatial estimation and forecasting can be applied to enhance decision support.

An example would be to combine the three NAFTA nations' censuses into a multi census GIS with data for 1990 and 2000. This would enable small area estimation of changes in dual metropolitan areas. It would further allow systematic projections of social and economic characteristics for cities and counties in border regions of the three nations. It could also compare binational border patterns between the U.S.-Mexico border and the U.S.-Canada border. These features could be build into a spatial decision support system, so they could be requested on-demand by users in all three countries.

Data Warehousing - key concepts

This section of the paper establishes the key features of data warehousing. In the next section, these key features are compared to the generic dual census GIS design in order to determine if a dual census GIS is a data warehouse.

A data warehouse is a very large data storage that emphasizes speed of access, multidimensionality, summarization, and flexibility/accessibility (Gray and Watson, 1998). It may be contrasted with a relational data-base (RDB) that accentuates data tables, relations, often complex queries, and inconsistency in user performance levels.

We utilize three aspects of data warehouses to characterize them: (1) data warehouse distinguishing features, (2) the OLAP model, especially as regards DSS aspects, and (3) steps in data flow. These are based on Codd (1970, 1995) and Gray and Watson (1998).

A first realm of characterization of data warehouses is distinguishing features, i.e. what distinguish them from other data or decision tools, such as RDBs, DSSs, and simulation models. Based on Gray and Watson (1998), we arrive at the following distinguishing features:

1. Performance tuned to decision-making users, rather than transaction users.
2. Data access is transparent i.e. the user doesn't know how or where the data are stored.
3. Time variant.
4. Non volatile
5. Multi-dimensional

Second, the data warehouses, in particular those applied to decision support, are characterized by the OLAP (On-Line Analytic Processing) model (Codd, 1970, 1995; Gray and Watson, 1998). OLAP consists of twelve major elements: (1) multidimensional view, (2) transparency to the user, (3) accessibility, (4) consistent reporting, (5) client-server architecture, (6) broad dimensionality, (7) dynamic sparse matrix handling, (8) multi-user support, (9) cross dimensional operations, (10) intuitive data manipulation, (11) flexible reporting, and (12) aggregation from unlimited dimensions.

This model emphasizes multiple dimensions, ease of use, consistency, and aggregation/summarizing. In the next section, the dual census GIS design will be compared against the Codd model.

Third, data warehouses are characterized by their steps of data flow. A data warehouse has a flow of data that moves from (a) operational data to (b) data cleaning to (c) data resident in a warehouse to (d) purging, summarizing, or archiving (Gray and Watson, 1998). This data flow reflects the aging of data from current/operational model to active data warehouse mode to disposal or archiving.

Comparison of Dual Census GIS to Data Warehousing

This section compares a dual census GIS with the three types of characterizations of data warehouses. The section determines whether or not a dual census GIS is a data warehouse and, in the case of OLAP elements, whether the dual census GIS as a data warehouse can be used for decision support.

Data Warehouse Distinguishing Features

The dual census GIS may be compared with the data warehouse distinguishing features. This is shown in Table 2.

Table 2. Comparison of a Dual Census GIS with Data Warehouse Distinguishing Features

	Data warehouse distinguishing feature	*Does the dual census GIS meet the data warehouse distinguishing feature?*
1	Performance tuned to decision-making users	YES. The dual census GIS is designed for decision makers as users, rather than transaction-based users.
2	Data access is transparent	YES. One of the benefits of the dual census GIS is that the data from two or more censuses can be easily, effortlessly, and transparently viewed and analyzed.
3	Time variant.	NO. This aspect is not critical to the dual census GIS, although it can be present to a limited extent. The limitation of time variation is the long periods between censuses, i.e. every 10 years.
4	Non volatile	YES. The dual census GIS is a decision making system that is not designed for volatile operational use.
5	Multi-dimensional	NO. The dual census GIS has limited dimensionality due to constraints mentioned above.

The dual census GIS mostly corresponds to the concept of data warehouse. The weaknesses occur in features emphasizing multi-dimensions and time variation. These weaknesses stem from the need to match characteristics, reducing dimensionality, and from the infrequency of census taking.

Codd Model

The dual census GIS design can be tested against each of the twelve OLAP elements. This is shown in Table 1.

Table 1. Comparison of a Dual Census GIS with Codd's OLAP Elements

	Codd OLAP Element	*Does the dual census GIS meet the OLAP element?*
1	Multidimensional view	NO. The dual census GIS has limited dimensions. What is unusual is the spatial dimension.
2	Transparent to the user	YES. The dual census GIS makes the inter-censal differences transparent to the user.
3	Accessible	YES. The dual census GIS increases accessibility because it overcomes a massive quantity of data.
4	Consistent reporting	YES. Reporting performance does not degrade as the size of the dual census GIS expands. At the same time, its updating is infrequent, since censuses are conducted every 10 years.
5	Client/server architecture	YES. Client server architecture is used.
6	Generic dimensionality	NO. There is not generic dimensionality, although the spatial dimension is included.
7	Dynamic sparse matrix handling	YES. The dual census GIS applies this element for missing data or adjusting data. For instance, in the U.S.-Mexico twin cities example, we had to impute missing data on the Mexican side.
8	Multi-user support	MAYBE. It depends on the scale and range of the dual census GIS.
9	Cross-dimensional operations	NO. The dual census GIS does not emphasize cross-dimensional operations. Rather, the spatial dimension is the crucial one.
10	Intuitive data manipulation	YES. In a dual census GIS, the results are best applied to decision making based on analysis and modeling. The user-friendly access to data manipulation is an important part of the dual census GIS.

	Codd OLAP Element	*Does the dual census GIS meet the OLAP element?*
11	Flexible reporting	YES. The dual census GIS can report through visual display, numeric tables, and statistical and other modeling outcomes.
12	Unlimited dimensions, aggregation	MAYBE. There is spatial aggregation and some attribute aggregation.. However, the dimensionality is more limited, so unlimited dimensions is not important.

It is clear that two thirds of Codd's OLAP elements are met by a dual census GIS. The ones that are not met generally stress multidimensional or cross-dimensional aspects. The dual census GIS has limited attribute dimensions due to the restrictions of having to achieve matches in attribute definitions. Also, the need for spatial referencing further constrains the number of dimensions, compared to a data warehouse.

Flow of Data

The dual census GIS shows lack of correspondence at the beginning and end of the data warehouse's flow of data. In particular, at the beginning of data flow, the data are not coming from an operational data-base, but rather from a one-time comprehensive census. At the end of data flow, the data are not purged or archived, but rather are maintained for the life span of the data warehouse. This reflects the much reduced and slower data updating process in a dual census GIS, versus a data warehouse, since censuses are generally held every ten years.

Further Application of the Data Warehouse as Part of an SDSS

If we assume the data warehouse definition is mostly met, then we need to ask, how can the dual census GIS be applied to modeling and decision making in the real world. GISs have been identified as a form of Spatial Decision Support System, or SDSS (Murphy, 1995). The Spatial Decision Support System includes a data-base module, model base, GIS module including analysis capabilities, and a user interface. This model could include a data-warehouse in place of the data-base module. The data warehouse can provide the SDSS with more rapid and more varied types of access to data. The reasons include the data warehouse's emphasis on accessing speed, efficiency, multidimensionality, flexibility, and ease of access.

The real power, however, comes from the model base working in conjunction with the data warehouse. In this way, the capabilities of modeling and analysis can be combined with multi-dimensional flexible data warehousing to give the maximal spatial decision support system (SDSS).

It is the OLAP features of the data warehouse that can make it an especially good component in this SDSS system.

On a practical basis, the dual census GIS with decision support can help government planners, analysts, and decision makers in two nations to work cooperatively with a common spatial data sets to make decisions, often ones involving both countries or related to relationships, interchanges and flows between the two countries. Also, a dual census GIS with longitudinal data, i.e. data from two or more census years, enables time variant models that can enrich the decision support capabilities with spatial forecasting, projections, and estimates.

Conclusions

This paper has reached its goal to present the general model of dual census GIS and compare it to the concept of a data warehouse. The research question is whether or not a dual census GIS corresponds to that of a data warehouse.

The principal finding is that the dual census GIS corresponds mostly to the concept of data warehouse. It shows good correspondence in the aspects of transparency, reduced volatility, accessibility, and flexibility.

On the other hand, it falls short mainly on concepts on multi-dimensionality and cross dimensionality, since the dual census GIS but puts emphasis on a single dimension, i.e. the spatial one. Constraints in attribute definitional matching reduce the dimensionality, as does the need for attributes to be spatially referenced. Another area of lack of correspondence is time variance. The dual census GIS does not emphasize time variance. One of the reasons is the infrequency of censuses. Where consistent census or other longitudinal data are available, that strengthens the dual census GIS, but it is less common.

The fairly good correspondence of dual census GIS to the data warehouse may be helpful in making some use of tools and concepts of data warehouses for these GISs. It is likely that very large-scale dual-census or multi-census GISs will also correspond to data warehouses; hence, for these sophisticated and large GISs, the data warehouse concept will be important since they have will have a very large volume of data (although the dimensionality and attribute sets may be similar). Another advantage of data warehousing design will be incorporation of OLAP functions, which provide greater ability for the dual census GIS to provide decision support features. This will strengthen the incorporation of the data warehouse into a SDSS to provide decision support to governments in decision making.

There are other advantages to the data warehousing concept for dual census GIS. The data warehouse can provide enhances query facility to users. This comes from the data access transparency and non-volatile features. Practically, these are important, since otherwise dual-census or multi-census users may be defeated in query attempts by the cumbersome obstacles to the dimensionality and by lack of access to longitudinal data.

Ability of the dual census GIS user to perform data scrubbing and aggregation are further benefits for the data warehousing approach. These would be formidably difficult say in the multi-census context, without the consistency and efficiencies of the data warehousing design.

The lack of complete correspondence reflects fundamental differences in public and private sector data and systems. For instance, dimensionality constraints, which are present for the dual census GIS, may be present in private sector applications but would be different. These basic differences could be explored more by controlled studies of government versus private sector data warehouses and by trying to find instances in private industry of GISs that combine two or more huge data sets, analogous to combining several censuses.

The skill set needed by designers of dual-census and multi-census GIS includes data-bases, GIS software including boundary analysis, and demographic/census understanding. On the other hand, the skill set of data warehousing designers and operators includes data-bases, data structures, decision support systems, and optimization. The close comparison of these two methodologies implies that data-warehousing skills may also be useful for multi-census GIS, and especially for larger multi-census GISs that involve data-warehousing like functions. As multi-census GIS becomes more refined, there might even be a merging of the two job descriptions into a common one in government settings.

Acknowledgment

The authors acknowledge the grant support for this project from the Ford Foundation. They also acknowledge the helpful comments of anonymous reviewers.

References

Barquin, R. C. and H. A. Edelstein (eds.). *Building, Using, and Managing the Data Warehouse*. Upper Saddle River, New Jersey: Prentice Hall, 1997..

Brackett, Michael H. *The Data Warehouse Challenge*. New York: John Wiley and Sons, 1996.

Codd, E.F. "A Relational Model of Data for Large Shared Data Banks." *Communications of the ACM* 13(6), 1970.

Codd, E.F. "Twelve Rules for On Line Analytic Processing." *Computerworld*, April 13, 1995.

Gray, P. and H. J. Watson. *Decision Support in the Data Warehouse*. Upper Saddle River, New Jersey: Prentice Hall, 1998.

Huxhold, William. *An Introduction to Geographic Information Systems*. New York: Oxford University Press, 1991.

Longley, P. A., M. F. Goodchild, D. J. Maguire, and D. W. Rhind (eds.). *Geographical Information Systems*. 2 Vols. 2nd edition. New York: John Wiley and Sons, 1999.

Murphy, L.D. 1995. "Geographic Information Systems: Are They Decision Support Systems?" *Proceedings of the 28th Annual Hawaii International Conference on System Sciences*, 1995, pp. 131-135.

Pick, J. B., N. Viswanathan, W. J. Hettrick, and E. Ellsworth. "Spatial and Cluster Analyses of Urban Patterns and Binational Commonalities in the Mexicali and Imperial County Twin Metropolitan Region." *Proceedings of the American Statistical Association, Section on Statistical Graphics*, 1999.

Pick, J. B., W. J. Hettrick, N. Viswanathan, and E. Ellsworth. "Intra-Censal Geographical Information Systems: Application to Binational Border Cities." *Proceedings of European Conference on Information Systems*, 2000, in press.

UN. "Global Urban Observatory: Monitoring Human Settlements with Urban Indicators." Draft Guide. Nairobi, Kenya: United Nations Centre for Human Settlements, 1997.