ENTITY-RELATIONSHIP APPROACH TO REGIONAL ECONOMIC DATA ANALYSIS

Roger J. Beck and Larry R. Medsker*

Introduction

Renewed interest in the formulation of policies that can enhance local area economic development has prompted requests for information about local economic trends. Questions are raised such as: (1) How does the current stressful situation in agriculture affect local economies? (2) Does the local economy simply react to national business cycles, or does the region perform better or worse than the nation because of the economic makeup of the region? (3) To what extent have public policies enhanced income and employment levels of local economies? (4) How are policies that favor business expansion likely to affect the natural resources of the region and the social and economic makeup of local communities? (5) Are the local communities equipped to deal with increased demands on public services if, in fact, business expansion occurs? and (6) What specific types of businesses are most likely to be attracted to a community with a given set of resource endowments? These are among the concerns of citizen groups, planners, and public officials that are operating in the policy environment called economic development.

Addressing questions such as these requires assistance from researchers who focus on specific components of the development process. Conceptual and empirical models of economic development are necessary to better understand the role of the resource mix and market conditions that can influence community economic growth (see, for example, Deavers and Brown, 1984). Secondly, public officials and agencies interested in implications of alternative policies need ready access to public data sources that provide information about the economic performance of the local economy. The present study focuses on an important source of data for economic development studies made available by the Bureau of Economic Analysis (BEA). That data source has utility for both research and public service needs, and we are giving particular attention to the latter.

*Respectively, Assistant Professor, Department of Agribusiness Economics, Southern Illinois University, Carbondale, Illinois and Associate Professor, Department of Computer and Information Sciences, Purdue School of Science, Indianapolis, Indiana.

This study is a joint effort by the Department of Agribusiness Economics at Southern Illinois University-Carbondale (SIU-C) and the Department of Computer and Information Sciences at the Purdue School of Science-Indianapolis. The results reported here come from the restructuring of magnetic tape files from BEA using two different approaches: (1) standard file manipulation to produce tables, and (2) entity-relationship data analysis followed by the formation of a relational database that satisfies normalization rules.

Description of the Data

The U.S. Bureau of Economic Analysis Regional Information System (REIS) makes available, to members of the user group, regional economic data produced at BEA. Members of the user group serve as distributors and field representatives for the data. As a member, the Agribusiness Economics Department at SIU-C interprets this responsibility to mean the provision of the BEA data in a readily accessible form to public and private entities throughout the southern Illinois region. This implies a need for flexibility in the manipulation of the data to meet specific requests.

The magnetic tape files from BEA consist of a COBOL print program, line title files, footnote files, and data files. For this study, we are interested in more robust techniques for providing processed subsets of these data, files which are made up of large (310-character) complex records. The format of the records (see figure 1) with a sequence of fields for each year's data, without access to a mainframe computer with tape drives is not convenient for exploratory examinations of trends of specific types of income. Disclosure codes, which are quite important to the analysis since a positive feature of the date is that they are available at the county level, are difficult to incorporate into the data manipulation process.

The specific data files on our BEA tapes cover (1) personal income by major sources and earnings by industrial source, (2) employment by type and broad industrial source, (3) transfer payments by a major source, and (4) BEA farm income and expenditures. The bulk of the county data used to estimate wage and salary disbursements is obtained from the ES-202 tabulations of the administrative records of the state's program of

unemployment insurance (U.S. Department of Commerce, 1983).

Current Approach

Our original approach to providing reports from the personal income file was to restructure the 310-character records into a format that was more convenient for data analyses. Statistical packages could then be used to reference a new variable (year), in conjunction with the Federal Information Processing Code (FIPS), for each political entity in the file. A particularly important type of analysis was to examine trends of a component of wage and salary income over time. Another modification was to place the disclosure code for a specific data item in a field adjacent to that data item. The product of this rearrangement of the data file, then, was a transposed data matrix.

This approach has the advantage that the transposed data set is more easily processed for trend analyses with a time variable using standard statistical packages. A problem, however, is the necessity of transposing the data matrix before data in this form can be made available to other users, who must then still separate out the data of particular interest to them. This can pose a major obstacle for local agency groups that may have little or no mainframe computing capability, and the process becomes a deterrent to local researchers and agencies who would otherwise like to have access to the data. Even though several reports have been generated from BEA data for Illinois (see, for example, Table 1) using the standard approach, that experience has motivated research, described in the next section, into better ways to structure the data for more flexible and convenient manipulation.

Experimental Approach

Another method for analyzing economic data is to make use of concepts from the entity-relationship (E-R) model (Chen, 1976; Chen, 1985; Chrisman, 1986) and the relational database approach (Date, 1983; Date, 1986; Ullman, 1982) to restructure the records into smaller ones that satisfy the rules of normalization theory. Our hypothesis is that the method described here allows not only a more practical way to work with economic data, but also provides a semantic structure that can give insight into the topic being studied.

In order to compare the alternative approaches, we need to be clear about the issues and problems that are involved in analyzing data. For example, the analytical method needs to provide:

* flexibility and convenience for data handling,
* the ability to perform a wide variety of analyses,
* convenience in updating the database as new information arises,
* ease of use, allowing the subject-matter specialist to concentrate on economic analysis instead of computer-related details,
* ease of access by the casual user to subsets of the data (e.g., a local user wants a portion of the data for a particular county).

Considerations such as these form the basis for comparing the alternative approaches.

Data Models

The entity-relationship model (Chen, 1976; Chen, 1985) was developed to provide a semantic framework for understanding data. At the beginning of a project, one has to identify the domain of data to be considered, identify a specific list of named data items, and clarify the meanings of data items and relationships between data items. At a practical level, a data model can provide techniques for diagraming the collection of data items and investigating their properties. An intentional and systematic approach promotes in-depth discussion of the data items early in a project before a commitment is made to a detailed implementation plan for manipulating the data.

From the subject-matter standpoint, the use of a semantic analysis raises interesting issues about the information at hand and about the objectives of the data analysis. This process commonly gives rise to new hypotheses to be studied and/or ideas for different or additional data to be collected. For the purpose of this paper, the convention shown in Figure 2 will be adopted for depicting a data collection. In this method, entities are defined as named, real-world objects or concepts about which data can be collected.

An entity is related to another entity in functional (1-to-1 or 1-to-many) or nonfunctional (many-to-many) relationships. The process of deriving an accurate diagram of the data being examined in a project involves serious consideration of the real meaning of the data and often leads to new and more accurate understandings. The results of a semantic data analysis must next be mapped into a data model that can be implemented using a database management system (DBMS). Although other models can be employed, the present project uses relational model concepts (Date, 1983;

Date, 1986) to further refine the organization of data items identified in the E-R analysis. In addition to providing a good theoretical approach, the relational model is important because it is receiving a lot of attention as the basis of new commercial DBMSs. A goal of relational DBMSs is to move data management closer to the technical level of the subject-matter expert.

The relational model has two features important for this paper: a relational structure for data and a set of relational operators for manipulating the data. The first aspect allows the definition of a particular collection of data items (called a relation) identified by the relation name, data-item names, and a key data-item name whose values uniquely identify specific values of the set of data-item names in that relation. With this data model, a relation (sometimes thought of as a table) is a random collection of tuples (or table rows), any one of which can be retrieved by specifying values for the relation name and key data-item. The particular assignment of data-items to a relation can be analyzed using, for example, normalization theory (Date, 1986) to see if the relation has any properties that might lead to problems in data update or deletion processes. More details on the relational model and normalization can be found in the references indicated above.

The other important feature of the relational model is that data manipulations can be expressed in logical forms closer to the level at which a subject-matter specialist thinks of the data. The mathematical notion of sets and operators is the basis for viewing data as relations upon which relational operators transform one or more tables of data into another table. Using these operators, the data analyst asks for different views of the database in order to look for relationships and trends and to produce reports.

Normalization theory provides rules for establishing the quality of a proposed set of relations. A relation satisfying particular conditions is said to be in a normal form (1NF through 5NF plus a Boyce-Codd Normal Form). The higher the normal form that obtains, the better the relation will be with respect to updates, insertions, and deletions of data values. For example, a relation in first normal form has only atomic data values -- i.e., no data-field entries can be a list of data values. A complete description of normalization is beyond the purpose of this paper but can be found in the references (e.g., Date, 1986).

Application of the Models

The theoretical approach outlined above was the basis of the alternate analysis of the BEA personal income data. First, the domain of possible data items, as provided on the tape from BEA, was inspected using the entity-relationship method in order to clarify the meaning of the relationships between the various data items. A goal was to find logical groupings of data that would be potential database relations. An additional benefit was an improved understanding of the data for further economic analyses that could be performed.

Figure 3 shows a portion of a larger network of entities, along with their attributes, that were identified from the E-R analysis. Diagrams such as this are derived after discussion, feedback, and as a better understanding of the data evolved.

The next step was to establish a tentative relation corresponding to each entity and to let the attributes from the E-R diagrams become names of the fields in the relations. Also, a field of each relation was chosen as the key--an attribute whose value uniquely identifies any row of actual data values in a relation. Finally, normalization theory was applied to see if any of the relations should be reorganized. The specifications for the final set of relations are given in Figure 4. As can be seen there, a characteristic of the relational model is some duplication of data; however, the overall quality and performance of a relational database is considered to be worth that cost.

The use of these relations is illustrated, along with sample data from the BEA data tape, in Figure 5. For the relation named Political_Division, any record is accessed by specifying a particular value of the key FIPS_Label; i.e., no two rows in that relation have the same values for FIPS_Label. This relation efficiently stores basic data associated with FIPS coding and can be referenced in database queries in order to access associated data in other relations. Likewise, combinations of fields--e.g., FIPS_No., Year--can be used to look up specific rows of data in the relation named Factor_Payments.

Figure 5 also contains an example of a query on the Factor_Payments relation to generate a new table of values that could be the object of statistical tests or that could become a table of a report. The query uses the syntax of the language QUEL, which is used in the INGRES DBMS. The languages associated with other systems will be somewhat different; however, the point here is that a query of arbitrary complexity can be written to construct any desired table from basic relations like those shown in Figure 5. Furthermore, the style of the queries is at a high, logical level that obviates the need for the user to be a programmer or to understand the details of the computer file system.

A preliminary implementation of the design described above has been carried out on a VAX 11/780

minicomputer and a Zenith (IBM compatible) microcomputer. The BEA data tape was read and transposed using a FORTRAN program that also divided the data into the relations described above. The relations were used as input to a DBMS, for which relational operators were constructed to produce data subsets. Similar tests were performed with dBASEIII using the data in relations captured onto microcomputer diskettes.

Discussion

Once data has been converted to the smaller relations, queries are easily constructed to produce a wide range of views of the data convenient for economic analysis and for generating requested tables. In the present project, the data conversion was a significant task requiring the kind of computing expertise the alternate method aims to eliminate. If an economic study is of sufficient length and complexity, the cost of coversion may still be justified. However, the desired situation would be to make the data available in relational form from the beginning.

The technique described here provides more options in viewing the data, allowing the economist to look at more data in an easier, more flexible way. Also, this type of analysis sometimes reveals gaps in the data, pointing to further data collection needed. As an example, examination of trends in income by county suggest that a more meaningful unit of analysis would be grouping of counties according to a degree of commonality in some key economic element. In our analyses (Beck and Herr, 1986), we have grouped the Illinois counties by their dependence on production agriculture. Then, we have collected additional information about the performance of agriculturally related business in those groupings of counties.

Another aspect of the technique is that new data is easier to incorporate into the established database: one either adds rows to appropriate relations or adds a new relation for each new concept or aspect added to the study. Beyond the utility for data handling, this approach guides the economist to a better understanding of the meaning of the objects about which data is being collected. Thus, E-R analyses can have impacts on economic development research, because the way in which one views the data, and the system under study, affects the framework of the research design.

Future work will be aimed at developing detailed metrics for comparing the performance of the alternate approaches. Data will be organized and analyzed in order to measure differences for a range of analysis types. Based on preliminary tests, the relational system has adequate response times to queries, indicating that a choice of techniques would rest on other requirements and considerations, as discussed above, rather than on physical data-handling performance.

One attractive feature of the relational approach is the modular nature of the data. An outcome of using this method could be that requested data could be delivered more efficiently by selecting the set of modules that meet the user's requirements. Thus, the amount of unneeded data that must be stored, shipped and processed would be minimized. Also, the modular data sets are amenable to analyses using the emerging software technologies such as integrated spreadsheets, DBMSs with intelligent interfaces, and other software for natural language and speech understanding. These developments, along with well-designed relational databases, can give an economist powerful analytical capabilities at a desktop personal workstation. For the local user groups, accessible data can enhance plant location studies, market research, and economic development planning efforts.

REFERENCES

Beck, Roger and William Herr. "Effects of Farm Sector Recession on Retail Sales and Nonfarm Income in Rural Illinois Counties." Paper presented at American Agricultural Economics Association Annual Meeting, Nevada, July 28-30, 1986.

Chen, P.P. "Entity-Relationship Model: Toward a Unified View of Data." *ACM Trans. on Database Systems.* 1(1976). 9-36.

Chen, P.P. "Database Design Based on Entity and Relationship." In S. Bing Yao (ed.). *Principles of Database Design.* Englewood Cliffs, N.J.: Prentice-Hall Inc., 1985.

Chrisman, C. and B. Beccue. "Entity-Relationship Model as a Tool for Data Analysis and Design."*Proc. 17th SIGCSE Tech. Symposium.* 18(1986). 8-14.

Date, C.J. *An Introduction to Databse Systems.* Reading, MA: Addison-Wesley Publishing Co., 1986.

Date, *Database: A Primer.* Reading, MA: Addison-Wesley Publishing Co., 1986. Washington, D.C.

Deavers, Kenneth L. and David L. Brown. "A New Agenda for Rural Policy in the 1980's."*Rural Development Perspectives.* 1(1984). U.S.

NGRES DBMS. Berkeley, CA:Relational Technology,Inc.

Ullman, J.D. *Principles of Database Systems.* Rockville, MD.: Computer Science Press, Inc., 1982.

U.S. Department of Commerce, Bureau of Economic Analysis. *Local Area Personal Income: Sources, Methods and Output Available from the BEA Economic Information System.* Washington, D.C.: 1983.

Table 1

Percentage change in income derived in farm-dependent sectors of Illinois counties grouped by relative importance of farm income to total income of county (1970-83).

Relative Importance of Farm Income in County	Retail		Service		Finance	
	1970-79	79-83	1970-79	79-83	1970-79	79-83
	(percentage change)					
More than 9.2%	55.2	7.1	81.5	50.9	187.9	21.1
4.40 to 9.19%	80.7	12.9	114.8	48.4	182.2	41.7
Less than 4.39%	87.8	16.3	133.4	52.5	143.1	43.1
State	86.0	15.7	131.4	52.2	145.8	42.7

Source: U.S. Department of Commerce, BEA, Local Area Personal Income, County Summary Tables.

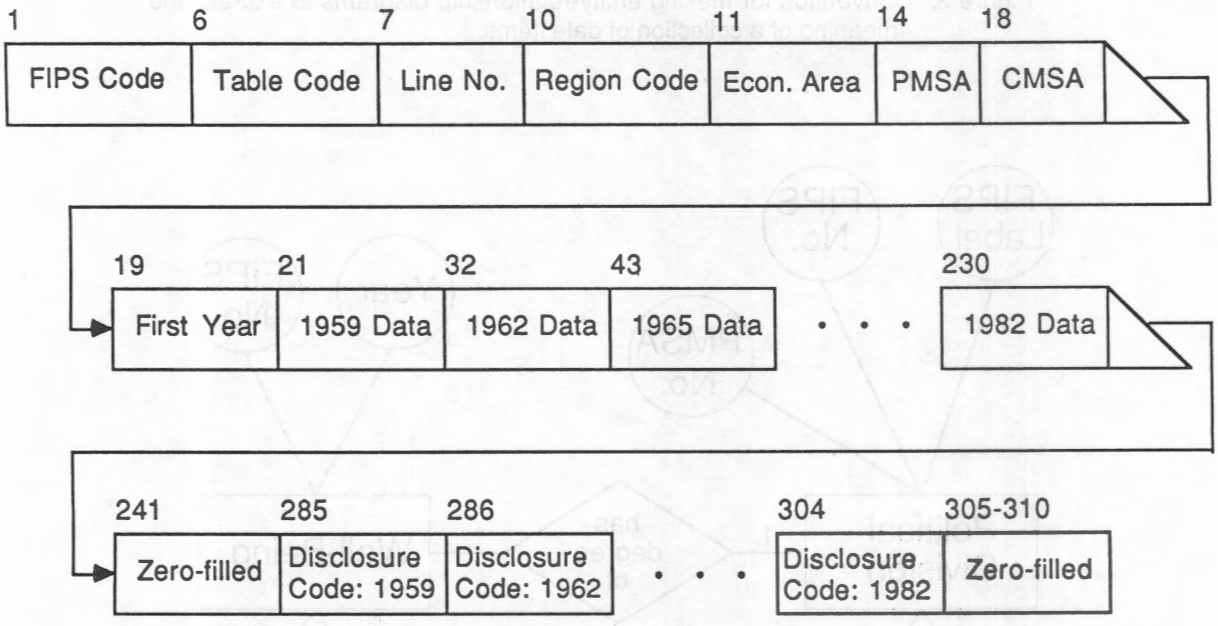

Figure 1. Record format for personal income data tapes. Line numbers specify the type of income source for which yearly values are found in fields 21-284.

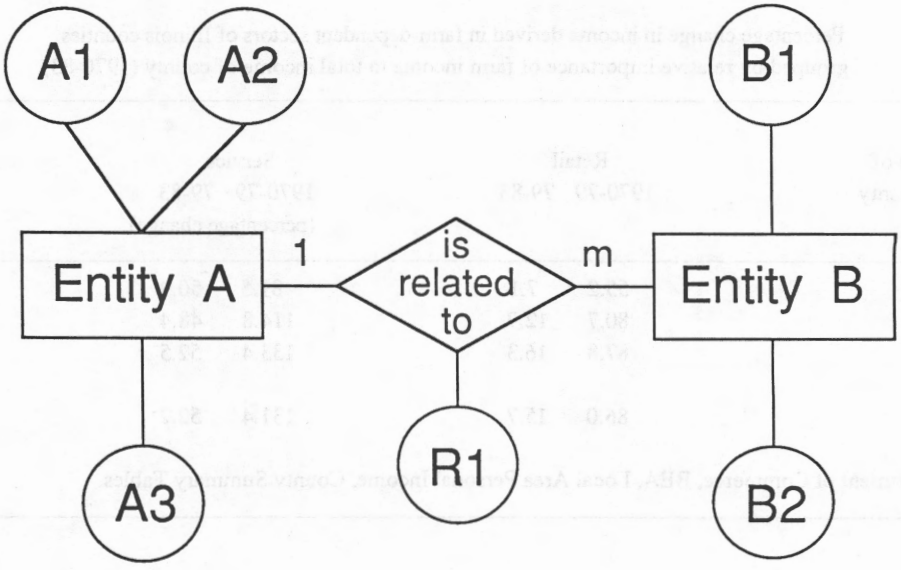

Figure 2. Convention for making entity-relationship diagrams to express the meaning of a collection of data items.

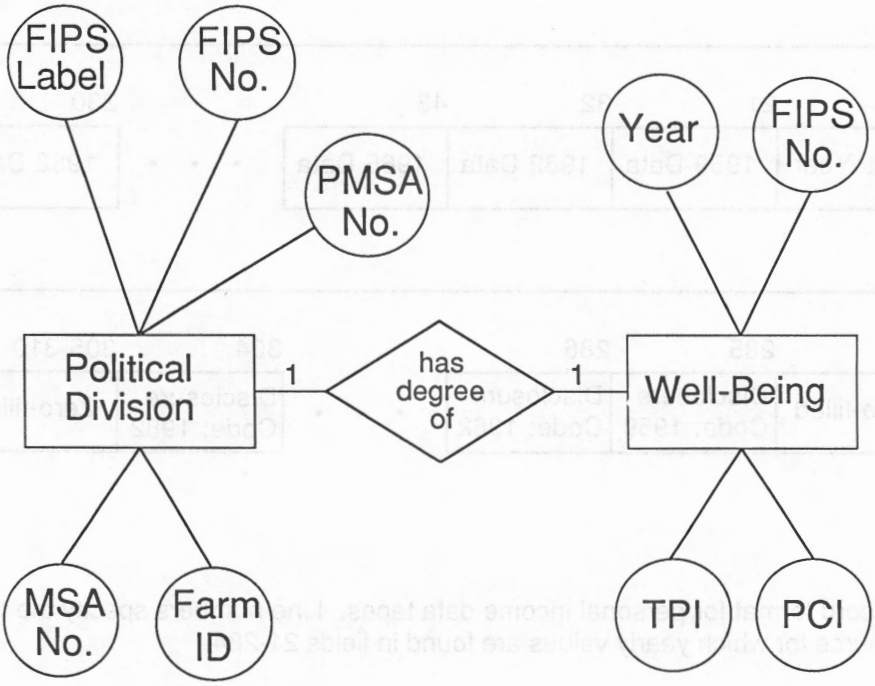

Figure 3. Sample Entity-Relationship diagram for a subset of the BEA personal income data.

Political_Division

FIPS Label	FIPS No.	PMSA No.	MSA No.	Farm ID

Demographics

FIPS No.	Year	Population	Over 65

Well-Being

FIPS No.	Year	TPI	PCI

Factor_Payments

FIPS No.	Year	Dividends	Transfer Payments	Sal. Distrib.	Total Prop. Inc.	Farm Prop. Inc.	Non-farm Prop. Inc.

Economic_Base

FIPS No.	Farm & Ag. Services	Mining	Construction	Manufacturing

Service_Sectors

FIPS No.	Transporation & Pub. Utilities	Wholesale Trade	Retail Trade	FIRE	Services

Public_Sector

FIPS No.	Fed - Civilian	Fed - Military	State & Local

Figure 4. Formats of relations derived for the BEA personal income data.

Political Division

FIPS_Label	FIPS_No	PMSA_No	MSA_No	Farm_ID
U.S. Total	00000	• • •	• • •	• • •
U.S. Metro.	00998	• • •	• • •	• • •
Illinois	17000	• • •	• • •	• • •
Rock Island Co.	17161	• • •	• • •	3
Alexander Co.	17003	• • •	• • •	3
Peoria	76120	6120	0	• • •
Chicago-Gary,IN-Lake Co.,WI	89914	899	14	• • •
•	•	•	•	•
•	•	•	•	•

Factor_Payments

FIPS_No	Year	Dividends	Transfer_ Payments	Sal_Distrib	Total_ Prop_Inc	Farm_ Prop_Inc	Non-farm_ Prop_Inc
00000	1959	484,445	27	2,573,157	472,678	104,139	37
00000	1962	592,020	34	2,949,597	495,470	119,110	38
•	•	•					
•	•	•					
17000	1980	208,774	15	754,200	53,601	6,057	5
•	•	•					
•	•	•					
17123	1975	162	0	238	235	195	0

Sample Query: Retrieve into Temp (FIPS_No, Year, Sal_Distrib, Total_Prop_Inc)
where (FIPS_No >= 17000) and (Year > 1970)

Result: **Temp**

FIPS_No	Year	Sal_Distrib	Total_Prop_Inc
17000	1980	754,200	53,601
•	•	•	•
•	•	•	•
17123	1975	238	235
•	•	•	•
•	•	•	•

Figure 5 Sample relations and query for BEA personal income data.