Spatial Data Integration for Low-Income Worker Accessibility Assessment: A Case Study of the Chicago Metropolitan Area

Piyushimita Thakuriah, Juan F. Ortega, P.S. Sriraj
Urban Transportation Center
University of Illinois at Chicago
412 S. Peoria, Suite 340
Chicago, IL 60607
vonu-pt@uic.edu, jorteg2@uic.edu, sriraj@uic.edu

ABSTRACT

The transformation of numerous and often disparate data sources into knowledge that supports critical decisions in a timely manner is essential in today's fast-growing market and digital government. Data integration is being increasingly recognized in the transportation sector as a valuable asset-forming activity aiding decision-making.

This paper will describe a spatial data integration application for a specific policy area: low-income worker accessibility in urban areas. This policy context is that entry-level job growth, which is appropriate for the skill level of welfare clients and low-income individuals, is mostly occurring far away from inner city neighborhoods or isolated pockets in the suburbs, where bulk of low-income and economically deprived families live. Many job-rich areas do not have affordable housing, including affordable rental housing, or there are implicit barriers to fair housing. Transit connections between job rich areas and where the bulk of the low-income and welfare recipients live (inner city neighborhood and isolated pockets in the suburbs) are limited and in some cases, non-existent. Vehicle ownership rates among such individuals are also very low. Most individuals in this category also do not have driver's licenses. Many entry-level jobs start at off-peak times. Job locations that are accessible during normal business hours are inaccessible during off-peak hours, since most transit services served peak hour transit markets.

There are three strategies of overcoming the spatial mismatch between jobs and home locations: (1) bring jobs closer to low-income neighborhoods (2) bring affordable housing solutions near employment generating areas and (3) provide transportation connections between jobs and housing. For effective and holistic solutions to the low-income worker accessibility solution, a database integrating information supporting all three strategies above is fundamental, in order to identify sub-markets that are likely to benefit most from a particular strategy.

To facilitate the planning and development of effective cross-sectoral strategies, a Geographic Information System that ties together data from the transportation, affordable housing and economic development sectors is currently under development. This paper will describe the architecture of this GIS and the data components of the system. A survey of urban geospatial data managers in relevant sectors in the Chicago area will drive the data requirements of the integrated system. Further, data quality issues of the integrated geospatial data will be described, focusing on coverage/scope, consistency, completeness, accuracy, timeliness and accessibility. Since this GIS is an integration of existing geospatial data, the scope of conflation tools and the impact of spatial aggregations in the disparate input databases will be documented. Since data specialists from the different sectors "speak different languages", the development of effective metadata is
also crucial. The paper will conclude with examples of the value-added information emerging from the integration process.

Key words: —data integration—data quality—GIS—low-income accessibility—transportation planning
INTRODUCTION

In recent years, the development of planning and evaluation tools for accessibility to economic opportunities by low-income has received increased attention. Entry-level job growth, which are appropriate for the skill level of welfare clients and low-income individuals, are mostly occurring far away from inner city neighborhoods or isolated pockets in the suburbs, where bulk of low-income and economically deprived families live. Many job-rich areas do not have affordable housing, including affordable rental housing, or there are implicit barriers to fair housing. Transit connections between job rich areas and where the bulk of the low-income and welfare recipients live (inner city neighborhood and isolated pockets in the suburbs) are limited and in some cases, non-existent. Vehicle ownership rates among such individuals are also very low. Most individuals in this category also do not own drivers licenses. Since many entry-level jobs which is the appropriate job pool for the educational and skill levels of individuals transitioning to work start at off-peak times, job locations that were accessible during normal business hours were inaccessible during off-peak hours, since most transit services served peak hour transit markets.

As transportation planners are increasingly called on to address such social considerations beyond historical mobility-based considerations in transportation planning, methods and tools that allow the study of these issues and which focus on accessibility concerns are needed (1). Motivated by these considerations, this paper describes the development process of one aspect of a Spatial Decision Support System (SDSS) that is currently undergoing development in the University of Illinois at Chicago (2). This SDSS is intended to help planners develop projects and policy makers take decisions about the Labor Market and Transportation (LMT) problem. The LMT problem is defined as “the synergistic strategies in the transportation, housing and economic development sectors such that spatial access to jobs is facilitated for low-income individuals”. The high-level architecture of the LMT-SDSS is described in a design document (2). The LMT-SDSS has two major components: a geospatial data component, and a decision-support component using the Analytical Hierarchy Process (AHP) (3). In this paper, we restrict our attention to the GIS component.

This paper describes the system development approach, the goals and objectives of the system and the criteria used a system design principles. First, the system described here is being developed by integrating the perspectives of key LMT stakeholders in the Chicago region from the transportation, housing, economic development and human services sectors with prior, ongoing work in the transportation sector on the same issues. Stakeholders provided their perspective on technical and institutional barriers to cross-sectoral data integration as well as their “requirements” on data content and media/format for information dissemination. The process of developing the GIS component of the LMT-SDSS, along with problem area objectives and Measures of Effectiveness (MOE’s) that serve as system design principles are described in Section 2. This section has several different subsections describing information content, system functions, system design. And second, it summarizes the process of developing the GIS portion of the LMT-SDSS is described, given the considerations described in the paper. Future work is described in Section 4.

GIS DEVELOPMENT PROCESS

To extend the learning that took place from the welfare to work and the EJ process to systematically include the needs and perspectives of the non-transportation stakeholders in the LMT problem area, we conducted a survey of urban geospatial data managers and policy makers from all four sectors (transportation, housing, economic development and human services). These
combined perspectives drove the information base for the LMT-SDSS. The survey was administered in a face-to-face interview setting. The survey included multi-choice, Likert-type scales as well as open-ended scenario based questions on three major issues: data/information content needed for LMT problem solutions, data dissemination methods/media and finally, perceived technical and institutional barriers. The methods used for data reduction included Soft Systems Methodologies (4) as well as an expert-opinion rating approach that is presented here.

**Goals, Objectives and Measures of Effectiveness**

The high-level goal that the GIS component addresses is to enhance information availability for the identification and implementation of LMT projects by closing the gap between existing, piece-meal information availability and desired integrated state of information availability. Information availability is enhanced when data from the different sectors are gathered together in a repository and warehoused in one place, available in one platform and “cross-linked” in some sense, information processing is seamless for data managers from all sectors and information dissemination tools and media support desirable knowledge-seeking.

This high-level goal was transcribed into two more targeted objectives, from which we developed Measures of Effectiveness (MOEs) which would ultimately drive the GIS design. The specific objectives that translate this goal into operational benchmarks guiding system development are:

1. Increase information and LMTP decision support;
2. Enhance communication and LMTP partnership development potential.

These objectives can be seen in Figure 1. The horizontal axis is the Information and Decision Support dimension and the vertical axis, the Communication and LMTP partnership dimension. The ideal system would enhance information availability such that the vertex (1,1) is reached. The MOE’s that would maximize both dimensions are also shown in the figure.

**FIGURE 1. GIS Development Principles - Objectives and Measures of Effectiveness**

**Objective 1:** Information and decision support refers to the ability of the GIS to provide reliable, credible and “good-quality” data to raise awareness of the problem, identify potential sub-populations and geographical areas in need of LMT solutions and develop potential courses of action that may benefit target populations and areas. To evaluate if the GIS increases this objective, it becomes necessary to examine the following MOE’s:

- Improve data content
- Enhance planning, program development & evaluation functions
- Increase data quality
- Improve information transparency
- Composite indices
- Enhance data dissemination tools
- Generate detailed and standardized metadata

Thakuriah, Ortega, Sriraj 4
• MOE1: Content/type of information that the system supports including core (inherited) data and composite (derived) data or indices

• MOE 2: Functional capability of the system (what the data are able to do) and

• MOE 3: Data quality.

**Objective 2:** Communication and partnership development potential of data, on the other hand, refers to the potential of the information to reach out to potential collaborative partners for developing and implementing LMT projects. As indicated in Section 1, development of partnerships is critical for the implementation of LMT projects and the problem awareness creation aspect of Objective 1 should be augmented by the strength of evidence of collaborative benefits. To meet this objective, it is useful to consider the following MOE’s:

• MOE 1: Transparency, intuitive and ease of understanding of data.

• MOE 2: Cross-linkage capabilities and composite indices (it will be seen later in the section that this MOE overlaps with MOE 1 of Objective 1).

• MOE 3: Attractive and easy to use information dissemination tools.

• MOE 4: Detailed and standardized metadata.

**Data Content**

There are many ways in which a typology of data in the GIS may be developed. For example, a typology may be developed simply on the basis of the sector: transportation data, housing data and so on. While this is an important perspective on the data, a more meaningful typology for the purposes of development are whether the data are core data or derived data. The contents of the GIS data warehouse as desired by the respondents include both core data and composite derived data. Table 1 shows the core (inherited) data that was desired whereas Table 2 shows the composite (derived) data that respondents would find useful. As described in Section 1, core data acquisition is a matter of importing existing data in the four sectors that are deemed relevant to the problem at hand and as such is a matter of data acquisition issues such as identifying the right contact person, signing (if necessary) data sharing agreements, checking data for consistency, completeness and other data quality indicators that will be identified in Section 4. On the other hand, composite data are those indices or estimates that require not only the core (raw) data but also a model of the underlying situation.

It quickly emerged from the survey that in all sectors, a measure of “spatial accessibility” (to jobs, social services and other destinations that allow the fulfillment of necessary activities) is highly desirable. It is important to note here what is actually a derived or composite measure in one sector may be perceived as raw data by another sector. A case in point, again, are accessibility measures. The transportation and regional science literature has seen accessibility measures going through an evolution from Hagerstrom’s time-space accessibility concept, the Hansen-type measure to the spatial interaction based measures that can be viewed as a “weighted index of travel cost and competition for the destination activity”. We found that when respondents unfamiliar to the transportation literature talked about accessibility, they usually conceptualized it as spatial proximity or in terms of travel times alone.
The survey results indicated that lack of transparency of composite or derived data in one sector by stakeholders in other sectors pervasive, but also that within the same sector, misconceptions regarding these differences persist across organizational levels (data manager versus administrative/executive staff). From the small survey that we did, it is not possible to generalize whether cross-sectoral differences in lack of transparency regarding data are greater or less than cross-organizational differences. However, we believe that the creation of necessary metadata and qualitative description of data would have high educational value.

TABLE 1. Core Data Needs Identified by Survey Respondents from Different Sectors

<table>
<thead>
<tr>
<th>Information Expectations</th>
<th>Transportation</th>
<th>Housing</th>
<th>Economic Development</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transportation</strong></td>
<td>Origin and destinations of trips</td>
<td>Location of affordable housing, food shelters, emergency housing shelters</td>
<td>Location of one-stop centers</td>
</tr>
<tr>
<td></td>
<td>Location of transportation services in relation to daycare, hospitals, jobs, one stop centers, schools and residences</td>
<td>Housing quality, availability, rents, costs and other housing characteristics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Transit operations</td>
<td>Crime rates, location of schools and other neighborhood quality data</td>
<td>Location of entry-level jobs</td>
</tr>
<tr>
<td><strong>Housing</strong></td>
<td>Data on transportation opportunities; Data on the proximity of transportation to housing locations</td>
<td>Location of affordable housing and household sociodemographic information</td>
<td>Employment statistics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Section 8 voucher recipients’ location.</td>
</tr>
<tr>
<td><strong>Economic Development</strong></td>
<td>Location of transit hub</td>
<td>Location of housing, affordable housing and housing characteristics</td>
<td>Employment and poverty statistics; Location of low wage jobs and industries.</td>
</tr>
</tbody>
</table>

The above table serves as a first step toward addressing the metadata needs and should be interpreted thusly. The three sectors of transportation, housing, and economic development are portrayed in the three rows, while across the table, in the columns are the data pertaining to each of these three sectors. Thus, reading from left to right in the first row of the table, we have the responses from survey participants affiliated to the transportation sector. The first column indicates the transportation data that they are used to working with and the remaining two columns portray the data needs from the housing and economic development sectors that transportation planners would like to know more about and possess in order to move toward a more holistic decision-making process. Since the survey respondents were a mix of people with varied background and training, it was difficult to elicit any further specificity about the data than that reflected in the table.

As a result of these issues, we used open-ended discussions and prior knowledge to separate out information desired in the GIS into core data (primary and secondary data sources that which we
can simply inherit from one or more sectors) and composite (derived) data (that which we need to model according to core data, some modeling rules and assumptions).

**TABLE 2. Composite Data Needs Identified by Survey Respondents from Different Sectors**

<table>
<thead>
<tr>
<th>Transportation</th>
<th>Housing</th>
<th>Economic Development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility of system to daycare, hospitals, jobs, one stop centers, schools and residences.</td>
<td>Location of affordable housing at the trip origin and transportation hubs.</td>
<td>Clusters of unemployment</td>
</tr>
<tr>
<td>Index to measure user perceptions on the reliability of the transportation system</td>
<td>Index on customer satisfaction.</td>
<td>Index of transportation barriers of low-income population.</td>
</tr>
<tr>
<td>Index of transportation barriers of low-income population.</td>
<td>Accessibility of affordable housing locations to jobs, employment and social services.</td>
<td>Jobs at trip destination</td>
</tr>
<tr>
<td>Measure of transportation as an index of opportunities.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accessibility of transportation to affordable housing locations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transportation connections and costs between origin and destination</td>
<td>Socio-demographic information on income, race, population of households.</td>
<td>Information of low-income population at an aggregate level or index.</td>
</tr>
<tr>
<td>Transportation system characteristics, operations and performance data</td>
<td>Neighborhood quality indices</td>
<td>Job growth rate indices</td>
</tr>
<tr>
<td>Socio-economic/demographic information on, traveler information</td>
<td>Gentrification rate indices</td>
<td></td>
</tr>
</tbody>
</table>

**Metadata and Data Quality**

As indicated earlier, the LMT-SDSS GIS inherits core raw data from different sectors, to the extent possible and derives composite indices on the basis of the core data. Hence, it is a large-scale data integration exercise. Nevertheless, the data quality assessment aspect is as important here as in the case of other data warehousing exercises. Since we are dealing with three totally different groups of stakeholders with wide differences in their view of data, approaches to dealing with data, data use and data histories, the creation of metadata is crucial. While metadata is often tersely described as “data about data”, in the current context, metadata consists of structured information that describes, explains, locates or otherwise makes it possible to retrieve, use or manage an information resource. The LMT-SDSS metadata describes the whole dataset: the data, its attributes and most importantly, its quality; they help to organize data collections, manage data resources and provide documentation for users.

Metadata does not include actual data sets and is instead a documentation of the content and quality of the data set it intends to describe. The Federal Geographic Data Committee (FGDC) created in 1990 has set the Content Standard for Digital Geospatial Metada (CSDGM). The FGDC has also been associated with the International Organization for Standardization Technical Committee 211 to develop a metadata standard.
In order to create the metadata for a data set, one has to understand the data set clearly and also decide on the metadata standard to adopt. The USGS has outlined a four-step process in the creation of metadata (http://geology.usgs.gov/tools/metadata). The four steps are (1) to assemble information about the data set, (2) to create a digital metadata file, (3) decide on the syntax of the metadata file, and (4) to ensure the accuracy of the metadata by checking the subject and content. The subject and content of the data sets can be kept under control by ensuring high standards of data quality. This exercise is a difficult enough task when dealing with just one data set or data sets from one sector. In the case of a cross-sectoral approach as adopted in this research, it becomes complex and this section focuses the attention on the difficulties embedded in working with diverse data sources and data sets.

The linking of separate databases gives rise to its own types of data quality issues (2). A key factor in the level of success of the data integration process with relational databases is the quality of the key/identification variables in the separate, input files. With spatial data, where disparate data are linked together on the basis of one or more spatial attributes, the level of “cleanliness” with the spatial identification data is critical.

### TABLE 3. Data Quality Dimensions and Criteria for the LMT-SDSS GIS.

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Data Quality Criteria</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contextual</td>
<td>Coverage/scope</td>
<td>Extent to which GIS repository covers sectors and attributes pertinent to LMT project development and collaboration development analysis</td>
</tr>
<tr>
<td></td>
<td>Completeness</td>
<td>Degree to which key identifier variables have values</td>
</tr>
<tr>
<td></td>
<td>Consistency</td>
<td>Degree to which separate, input databases from different sectors are spatially and temporally consistent.</td>
</tr>
<tr>
<td></td>
<td>New Value Created</td>
<td>Extent to which linked repository adds new information for LMT-SDSS objectives over and above that possible from analysis of separate, input databases</td>
</tr>
<tr>
<td></td>
<td>Timeliness</td>
<td>Degree to which specified data values are up to date and reflects the temporal updating cycle of input datasets</td>
</tr>
<tr>
<td>Intrinsic</td>
<td>Accuracy</td>
<td>Extent to which key data elements are error-free and measures the intrinsic properties of attributes</td>
</tr>
<tr>
<td>Representational</td>
<td>Documentation &amp;</td>
<td>Degree to which print and online documentation of input databases are available in a timely, accurate and readable format for free or at nominal cost.</td>
</tr>
<tr>
<td></td>
<td>metadata</td>
<td></td>
</tr>
<tr>
<td>Accessibility</td>
<td>Ease of use and</td>
<td>Degree to which LMT-SDSS information uses input databases that are accessible easily, without written contracts, data sharing agreements and special permission and also the extent to which users can use information from the system without a great deal of prior knowledge and in a format/media that can be used easily using standard equipment and software</td>
</tr>
<tr>
<td></td>
<td>transparency</td>
<td></td>
</tr>
</tbody>
</table>

We identify four core dimensions of data quality consisting of eight measurable criteria (Data Quality Criteria) that should be satisfied by the GIS. These are shown in the left-most column of
Table 3. The second column of the table shows the DQC’s that operationalizes the core dimensions.

While researchers have accepted no single definition of data quality, there is agreement that data accuracy, currency, and completeness are important areas of concern. Based on prior work that one of the authors of this paper had conducted with data linkage and data quality in the transportation safety sector, we describe here a framework for data quality assessment for the LMT-SDSS GIS.

In order to integrate the data from Table 1, the following steps are suggested: 1. Identify data sources with the relevant data. 2. Decide on a spatial aggregation unit, for instance, census tracts or quarter sections. 3. If datasets are GIS coverages, make sure the projection is compatible with other datasets. 4. If datasets are in any text format such us excel or dbase, decide on what software should be use to clean and join datasets, our suggestion is to leave the datasets independent and join the tables. 5. Create the necessary composite indices according to the planning needs, for example accessibility and/or mobility measures. These same steps will be followed in the implementation of the case study.

It is clear from the discussion on data quality that building the metadata and ensuring data quality are tied together in more ways than one and are intrinsic to the successful development of a prototype. The prototype design envisioned for this research involves the use of the ArcGIS suite of software.

The system design for the prototype is as follows: It is embedded on Linux, and Oracle platform.

The data are imported into Oracle and the metadata are extracted with the help of ArcGIS ArcCataglog. These metadata documents are stored in ArcSDE and ArcIMS hosts the data application. In other words, ArcIMS enables the dynamic querying capabilities to the end-user at a remote location (ESRI, 2002).

**SUMMARY OF SYSTEM DESIGN PROCESS**

The data architecture of this system is unique; nevertheless it may be implemented using different combinations of software and hardware. It is necessary to incorporate (1) a database management tool, (2) a geographic information software, (3) an interface for the former two, and (4) a dissemination processor. The system designed by the Urban Transportation Center is the result of the human and technical resources available. Figure 2 shows the data sources and steps necessary for the design of the system. Similarly, it shows the two possible outputs of the system: (1) a GIS tool for displaying results and (2) an instrument to perform scenario analysis to tackle problem situations and develop policy guidelines.

For the implementation of the proposed system, a data warehouse is being deployed in a Linux Intel server (Linux 9.3), in Oracle 9i Database Release (9.2.0.1.0). The system has two data dissemination aspects: off-line/static and online/dynamic. The purpose of this dissemination falls into three different levels of end-uses: (1) long-term strategic planning decisions, (2) short-term policy analyses, and (3) project coordination purposes.

The off-line/static dissemination pertains to the derived data developed in-house at the Urban Transportation Center (UTC) and will be in the form of graphs, static maps, and tables. Data from the Consumer Expenditure Survey (CES), the Current Population Survey (CPS), and the
Decennial Census are included as part of this dissemination effort. For example, the Current Population Survey releases information about various aspects such as the labor force, industry, work times, etc. at different times of the year. The raw data gives information, say, about the work start times of a sample of the population. This data can be enriched by developing a model to predict job start times, if the information about an individual’s area of work (industry), age, education level, gender, etc. are known. This information in turn can help transportation/transit planners to plan for the future based on the travel patterns of people in a region.

The on-line/dynamic data pertain to the data from the three sectors and will be mounted on the ARC-GIS tools for dissemination. For example, data from the census can be queried on-line to extract a wealth of information pertaining to a range of transportation related issues. This can also be combined with data from a region’s transit provider/network of roads to enrich this information.

![Figure 2. System Design Process](image-url)
CONCLUSIONS AND RECOMMENDATIONS

It is a titanic mission to assemble all the data from the various sectors, therefore this data architecture will be built by the incremental inclusion of more variables, indices and models that will lead to the creation of different applications and eventually to the creation of Participatory-GIS. As a subset of the proposed SDSS, general public will be able to identify information such as housing, daycare, job locations, transportation network and transit. Data will be made available at any feasible level of resolution that the researcher might seek. The difficulty typically arises in either collecting the data, or in integrating different data sets together to enrich them. The barriers to data collection are addressed next followed by issues associated with data integration.

Organizational barriers: (1) Data analysts and data managers are almost possessive about the derived data that they develop and are more often than not reluctant in sharing their efforts with people outside their own department/agency. (2) At a higher level is the nature of the data, i.e., whether it affects individuals directly or reveals the identity of the individuals that lead organizations to guard them vehemently.

Geographical barriers: (1) Data are typically collected at various levels of resolution depending upon the objective of the research that they are being collected for. Thus, when data are shared outside of their own research environment, the question of aggregation/disaggregation becomes a big issue that presents as a barrier to integration of data.

Temporal barriers: Similar to the geographical barriers are the temporal barriers, which can be both microscopic and macroscopic in nature. They can be pertaining to the time of day/month that the data were collected or there can be issues about integrating data from different time periods. A typical example can be comparing the decennial census data from the year 2000 with the Chicago Transit Authority’s (CTA) ridership information from, say, 1995.

Technological barriers: While this can be overcome with the right training and investment, organizations should ensure that it is not overlooked. It is very essential to ensure that the appropriate training methods are made available, and monies are set-aside in order for individuals to cross the technological divide.

The organizational, geographic, temporal, and technological barriers are significant, as are the issues associated with data integration. The most important aspect of data integration in such efforts, as discussed in this paper, is to ensure that the quality of data are not compromised to satisfy short term goals/needs of decision-makers. It is in this respect that developers of such systems should understand the marriage between data quality and metadata and how they feed off of each other.
REFERENCES


