UCI-ITS-AS-WP-98-3

Incorporating Yellow-Page Databases in GIS-Based Transportation Models

UCI-ITS-AS-WP-98-3

Ming S. Lee
Michael G. McNally

Department of Civil and Environmental Engineering and
Institute of Transportation Studies
University of California, Irvine, mmcnally@uci.edu

August 1988

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-3600, U.S.A.
http://www.its.uci.edu

Incorporating Yellow-Page Databases in GIS-Based Transportation Models

Ming-Sheng Lee[1] and Michael G. McNally[2], Member, ASCE

Abstract

A systematic approach is developed to transform data in the existing yellow-page databases to a point-based GIS database on activity supply. Such a database is needed for an activity-based travel forecasting system and for disaggregate accessibility analysis. First, the linkage between activity types and business types is established. According to this lookup relationship, businesses and services associated with certain activity types can be selected. These records are then geocoded by address-matching in a GIS and the locations supplying those activities are pinned down. Technical issues, such as difficulty in linking businesses to activities, long term projection, and address-matching, are discussed and potential solutions are provided. Finally, issues that need to be addressed when attempting to develop an activity-based forecasting system are examined from the perspective of activity supply.

Introduction

Deficiencies of the current travel forecasting procedure (i.e., the four step procedure) in meeting requirements of federal legislation are well documented (Karash and Schweiger, 1994). Among these, the inability to project travel demand for air quality estimation is a fatal drawback. For example, vehicle miles of travel (VMT) by hour of the day by grid square (i.e., 2-km or 5-km-square grids) is

[1] Graduate Student Researcher, Institute of Transportation Studies, Department of Civil and Environmental Engineering, University of California-Irvine, Irvine, CA 92697-3600

[2] Associate Professor of Civil Engineering, Institute of Transportation Studies, Department of Civil and Environmental Engineering, University of California-Irvine, Irvine, CA 92697-3600

required for estimation of ozone emissions. Typical geographical units of traffic analysis zones (TAZ) are Census Tracts, which are predominantly irregular polygons. They can not be easily adapted, without loss of precision, to facilitate partitioning regional VMT at the desired resolution. Typical time-slices of the conventional forecasts, peak and off-peak hours, also fail to meet the requirement. In addition to detailed VMT estimates, the estimation of the number of trips made in cold start mode is also critical for determining emissions. The modeling of trip chaining is especially important in this respect because it affects the number of vehicle trips in cold start mode. Limited by the assumption that there is no intermediate stop in a trip, current trip distribution methodology locked on the TAZ system is incapable of modeling multi-stop trips. To model trip chaining behavior, the potential locations for stops of various purposes need to be identified within each TAZ. The zone-based approach also fails in the evaluation of certain transportation control measures (TCMs) (Stopher, 1993), such as employer-based trip reduction plans, trip-reduction ordinances, and employer-sponsored flexible work hour programs. To estimate the impact of these measures requires that the projection of work trip destinations be spatially allocated to where the relevant participants are.

Many researchers in the field of travel demand forecasting believe that the activity-based approach has the highest potential of offering forecasts with the spatial and temporal fidelity required by the federal legislation (McNally, 1997; Spear, 1994). The major tenet of this approach is that travel demand is derived from the needs for activity participation. When individual travel demand, for a given day, is realized according to the list of activities pursued and times spent in each activity, the resultant trip-ends are allocated spatially at the activity locations and temporally on the time horizon of the day. Hence, the activity-based approach is capable of producing point-to-point (i.e., activity to activity) forecasts. Although the activity-based forecasting system is still under development, Goulias (1997) identified data needs for such a system and divided these data into demand side and supply side. Data related to household background, logitudinal and geographic information on activity participation, and available household resources (i.e., means for travel and communication) are on the demand side. On the supply side, information on activity and networking opportunities (i.e., the supply of transportation and telecommunication) needs to be collected. The 1994 Portland activity/travel survey has collected most of the required data on the demand side by recording activities performed by each respondent over a two day period. When implemented with a facility for synthetic population generation (see Beckman et al., 1996), the prerequisites for micro level travel forecasting on the demand side are almost completed. However, the counterpart of the Portland survey on the supply side does not exist yet. Current zone-based land-use databases are too coarsely defined to pinpoint relevant activity locations.

There are other needs for a better activity supply data in regional planning. It has been argued that the zone-based approach applied in the past is not suitable for measuring individual accessibility, since the same level of accessibility is indicated

for people in the same zone (Pirie, 1979). Planners are currently engaged in an effort to devise accessibility indicators that reflect impact of transportation/land-use policies on fulfillment of individual activity needs; Handy and Niemeier (1997) showed that a database containing the geographic coverage of various commercial activities can be very useful for this purpose. However, their database was defined at the community-level; integrated, metropolitan-level databases are rarely available.

The lack of suitable data on activity supply reveals both the urgent need for enhancement of the current digital databases and the institutional issues regarding the relationship between the public sector, where most resources are held, and the private sector, where actual model development occurs. Existing yellow-page databases provide a potential solution to this dilemma. These commercially available databases represent the most thorough collection of data on supply of services. Via techniques of geocoding in a Geographic Information System (GIS), yellow-page listings can be converted to a point layer in which each point represents a potential activity location. The purpose of this paper is to propose a scheme for development of such an activity supply database.

Yellow-Page Databases

The term yellow-page commonly refers to books with listings of business phone numbers. Modern information technologies have extended the format of yellow-pages from exclusively hard copies to digital databases. A compact disc (CD) is ample enough to store virtually every business in the United States. A digital yellow page database may contain the following data items for each business: (1) Business name, (2) Phone number, (3) Street address, (4) County, city, state, and ZIP code, (5) Latitude and longitude of the location, (6) The Standard Industrial Classification (SIC) code, (7) Number of employees, (8) Annual sales volume, and (9) Year established. These databases usually come with interactive search engines that allow users to search for only records of a given street, city, state, ZIP code, area code, and/or SIC code. For example, officials in a city planning authority may search for records only in their city. The search results can be exported in all kinds of digital file formats (e.g., comma delimited text, dbase). for post processing. Although vendors offer products that are available over-the-counter, their use is not recommended for the scheme proposed here. Since these products are limited in both data items and search ability, customized databases are preferred. Such databases are available by contacting the vendors at a special cost per listing.

Activity Types vs. Business Types

To transform business listings to activity system supply, the linkage between business types and activity types must be established. The business type of each listing can be identified by its SIC code. This coding system was developed by the

Office of Management and Budget for use in the classification of organizations by types of activity in which they are engaged ("SIC Manual", 1987). The structure of SIC can be illustrated by the following example, which shows the hierarchy of retail businesses in descending order of aggregation:

Division, G: "Retail Trade",
Major Group, 59: "Miscellaneous retail",
Industry Group Number, 594: "Miscellaneous shopping goods stores",
Industry Number, 5941: "Sporting goods and bicycle shops".

By referring to the codes, listings associated with a specific activity can be identified and selected. For example, out-of-home meal activities can be supplied by businesses with major group code 58, "Eating and Drinking Places". Grocery shopping activities can take place at those with code 54, "Food Stores". Table 1 establishes the connection between activity types and SIC codes. This activity classification scheme is used in the Portland survey (Lawton, 1997). For each activity type, the actual destinations reported by the respondents are examined to determine what businesses supply that activity. It has to be noted that the right hand column of Table 1 contains mostly the two digit major group codes, because they cover most businesses patronized by the respondents. This does not imply that all businesses in these groups are qualified for that activity. Some businesses are actually irrelevant. Further processing should be performed at the lower hierarchy to screen out inappropriate businesses. For example, if the database is used for surface transportation planning, businesses or services that do not require traversing space for patronage should be excluded. "Non-store retailers" (code 596) thus should be excluded from the "Miscellaneous retail" (code 59) for representing shopping supply.

Converting listings to a GIS point layer

Once the businesses representing activity supply are identified and selected, they need to be geocoded in a GIS to determine the locations. Some database vendors already performed this task and included the longitude and latitude of each business in the data CD. However, users should be warned that certain businesses may be geocoded to the centroids of the ZIP code areas in these "geocode-ready" databases. To control for the spatial resolution, it is not recommended to use the coordinates generated by the vendors. Geocoding via address-matching is the preferable approach. The address-matching program reads the address of a listing and automatically finds the matching street segment in the reference street network (e.g., U. S. Census's TIGER/Line file). A point is then created, in a separate GIS layer, right next to this segment to represent the location of this business. Since the locations are anchored on a street network representation, the network travel impedance between two activities can be realistically measured. This feature enables the precise measurement of individual accessibility that can not be achieved

by aggregate zoning system (Kwan, 1998). After all points are matched, the resultant GIS layer is a point-based representation of the activity system (Figure1).

Table 1. Linkage between activity types and business types

ACTIVITY TYPES	BUSINESS TYPES (SIC CODES)
Meals	Eating and drinking places (58)
Work	N/A[3]
Work-Related	N/A
Shopping (General)	General merchandise stores (53) Food stores (54) Miscellaneous retail (59) Video tape rental (784)
Shopping (Major)	Building materials and garden supplies (52) General merchandise stores (53) Automotive dealers and service stations (55) Furniture and homefurnishings stores (57) Miscellaneous retail (59) Auto repair, services, and parking (75)
Personal services	Personal services (72) Auto repair, services, and parking (75) Child day care services (835)
Professional services	Auto repair, services, and parking (75) Legal services (81) Accounting, auditing, and bookkeeping (872) Miscellaneous personal services (729) Veterinarians (0742)
Household or personal business	Depository institutions (60) Post Offices (4311)
Household maintenance	Home[4]
Household obligations	Home
Casual entertaining	Amusement and recreation services (79) Eating and Drinking places (58)
Formal Entertaining	Amusement and recreation services (79) Eating and Drinking places (58)
Medical care	Health services (80)
School	Educational services (82)
Culture	Amusement and recreation services (79) Educational services (82)
Visiting	N/A
Out of area travel	N/A

[3] See Technical Issues for explanation

[4] These activities were mostly performed at home locations.

Table 1 (Continued)

ACTIVITY TYPES	BUSINESS TYPES (SIC CODES)
Religion/Civil Services	Religious organization (866) Civic and social associations (864) Educational services (82)
Civic	N/A
Volunteer work	N/A
Amusements (at home)	Home
Amusements (Out of home)	Amusement and recreation services (79) Motion picture theaters (783) Eating and Drinking places (58)
Hobbies	Home
Exercise/Athletics	Physical fitness facilities (7991) Amusement and recreation services (79) Educational services (82)
Rest and Relaxation	Home
Spectator Athletic Events	Educational services (82) Commercial sports (794)
Pick-Up-/Drop-Off passengers	Educational services (82) Child day care services (835)
Incidental travel	Home
Tag along travel	N/A

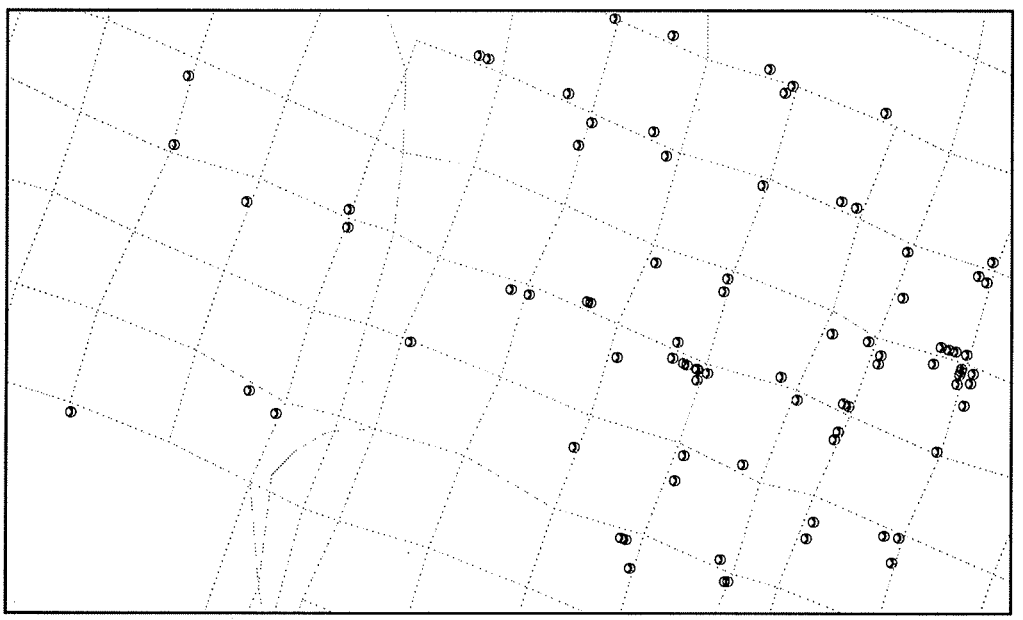

Figure 1. Point-based activity system layer

Technical Issues

Issues related to activity-business linkage

It can be seen in Table 1 that several activities were not linked to any businesses or services. Except for activities that happen exclusively at home and out-of-area travel, these activities present various dilemmas when identification of their potential locations are needed. The problems can be categorized into four types. First, the proposed approach may not be suitable for identifying the potential locations for work activities. Since all businesses and services are qualified candidates, the massive number of listings will impose a huge overhead on data processing. Even if the task of data processing was managed, the highly concentrated points would make the layer look similar to a zone-based layer. Hence, such a database may not necessarily be different from a zone-based land use layer when used to represent work opportunities. The use of commercial and industrial zones with appropriate disaggregation should be considered. Second, work-related, and tag along trips were performed at various destinations. It is hard to find distinct business groups to cover the potential locations. Since the occurrence of these two activities are spontaneous in nature, probabilistic simulation may be the only way to model the demand and supply of these two activities. Third, volunteer work and civic activities were also performed at various destinations. However, a certain portion (i.e., about 50~60 %) of these activities took place at churches, schools, or community centers. To use them as potential destinations should provide reasonable accuracy with respect to model outputs. Finally, there is no establishment providing services to visiting activities. The destinations of this activity are exclusively residences of friends or relatives. Although residential zones can be used as the surrogates of visiting destinations, the broad coverage of these zones in an urban area makes the pinpointing task an elusive one.

Issues related to longitudinal evolution

Land use models have long been a weak component in urban/transportation planning (Weiner and Ducca, 1996). It is not clear at this point as to how the activity system evolves at the micro level. To project the long term evolution for the proposed database presents a new challenge to the activity-based modelers. Some yellow-page databases include the establishing year of each business that may be a feasible point to start the effort. Without an evolutionary engine, the proposed point layer is suitable for evaluation of short run TCMs that do not require the projection of changes in the activity system. Incorporation with the Advanced Traveler Information System (ATIS) is another potential usage of this database. In addition to route guidance, drivers can acquire information to substitute locations of the same (or a different) activity type when situations dictate.

Issues related to address-matching

Although geocoding via address-matching is generally considered by GIS professionals the most cost-effective way of obtaining locational information (Drummond, 1995), it has certain limitations. When the address to be matched can not be found in the reference street network, the matching fails. The failure can be caused by errors in either the reference streets or the target databases. At the reference's end, typical errors include out-dated or missing streets and number ranges. Due to the popularity of GIS applications, there are many geographic data companies and metropolitan planning organizations offering enhanced street databases. These databases update the TIGER/Line file and add missing data to it. Match rate can be increased by adopting such enhanced street databases as references. On the target's end, addresses in a yellow-page are usually listed in the street format or non-street format. The former conforms to the standard "number, street, city, state, ZIP code" format. Listings in this group generally can be matched with an enhanced reference. Non-street addresses refer to those listed as Post Office boxes or special landmarks such as malls, plazas, and complexes. These addresses would not be included in any street databases. As mentioned previously, businesses or services that do not invite travel should be excluded. Thus, address-matching for listings with Post Office boxes is not necessary. Listings associated with a landmark usually cluster around intersections. Locations like this can be identified by a phone call to one of these businesses or by referring to the prevalent Internet map browsers. Once the location of a landmark is pinned down, all listings at this location can be matched at once. For example, a dummy street link can be created at this location as the reference. The number of establishments and the name of this landmark can be coded as the dummy street range and name, such as 1-100, the Mall. By re-running the address-matching program, businesses No. 1 to No. 100 in "the Mall" can be geocoded at once.

It has to be noted that there are companies providing geocoding services. If funding is available, geocoding of listings can be accomplished by the professionals.

A practical example

The feasibility of the proposed approach is demonstrated by Lee and Goulias (1997). A GIS database containing all shopping opportunities in State College, PA was developed for analysis of accessibility. This database enabled the precise measurement of individual accessibility to shopping, because the GIS databases, including the transportation network and shopping activity supply, were organized as a "snapshot" of the real world. The results of the analysis showed that the proposed accessibility indicators are significant explanatory variables of the weekday shopping trips. Although problems with address-matching did happen, addresses updated by the township authorities and mail-in surveys helped identify locations of unmatched businesses (Lee, 1996).

Summary and Conclusions

From the perspective of activity supply, there are several issues need to be taken into account when developing an activity-based forecasting system. First, it may be necessary to aggregate some activity types defined in the Portland scheme (see Table 1). For example, the difference between "Personal services" and "Professional services" is not clear. Some respondents reported activities taken place at automotive services as "Professional services" and others reported them as "Personal services". This is also true for "Casual entertaining" and "Formal entertaining". Any attempt to deriving household activity patterns from the Portland database should notice the existence of such confusion. Since allocating trip-ends to activity locations is the major tenet of activity-based approach, aggregating activities on the basis of potential supply locations should be one way to increase modeling tractability. Second, the potential locations supplying certain activities are hard to target. As discussed previously, work, work-related, tag along trips, civic, volunteer work, and visiting require special consideration. Finally, many activities can be supplied either out-of-home or in-home. Meals, casual entertaining, formal entertaining, and exercise/athletics are some of them. The modeling of substitution between out-of-home activities and in-home activities is a subject that must be addressed.

Acknowledgements

This research was supported in part by the U. S. Department of Transportation and the California Department of Transportation through grants to the University of California Transportation Center. This paper continued the work done by the first author at the Penn State University. The authors thank Konstadinos Goulias of Penn State University and Mei-Po Kwan of Ohio State University for the inspiration and comments.

References

Beckman, R.J., Baggerly, K.A., and McKay, M.D. (1996). "Creating synthetic baseline populations." *Transportation Research A*, Vol. 30, No. 6, 415-429.

Goulias, K. G. (1997). "Activity-based travel forecasting: What are some issues?" *Proceedings, Activity-Based Travel Forecasting Conference*, Travel Model Improvement Program, 37-49.

Handy, S. L., and Niemeier (1997). "Measuring accessibility: An exploration of issues and alternatives." *Environment and Planning A*, 29, 1175-1194.

Karash, K. H., and Schweiger, C. (1994). "Identification of transportation planning requirements in federal legislation." *Report prepared for John A. Volpe National Transportation Systems Center*, DOT-T-94-21.

Kwan, M.-P. (1998). "Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework." *Geographical Analysis*, (Forthcoming).

Lawton, T. K. (1997). "Activity and time use data for activity-based forecasting." *Proceedings, Activity-Based Travel Forecasting Conference*, Travel Model Improvement Program, 103-117.

Lee, M. (1996). "Analysis of accessibility and travel behavior using GIS." *Master Thesis*, The Pennsylvania State University.

Lee, M., and Goulias, K. G. (1997). "Accessibility indicators for transportation planning using GIS." *Paper Presented at the 76th Annual Transportation Research Board Meeting*, Washington D. C., January 12-16,1997.

McNally, M. G. (1997). "An activity-based microsimulation model for travel demand forecasting." *Activity-based approaches to travel analysis*, D.F. Ettema and H.J.P Timmermans, eds., Elsevier Science, New York, New York.

Pirie G.H. (1979). "Measuring accessibility: A Review and Proposal." *Environment and Planning A*, Vol. 11, 299-312.

"Standard Industrial Classification Manual." (1987). Office of Management and Budget, Executive Office of the President, Washington, D. C.

Stopher, P. R. (1993). "Deficiencies of travel-forecasting methods relative to mobile emissions." *Journal of Transportation Engineering*, Vol. 119, No. 5, 723-741.

Spear, B.D. (1994). "New approaches to travel forecasting models: A synthesis of four research proposals." *Report Prepared for U.S. Department of Transportation*, DOT-T-94-15.

Weiner, E., and Ducca, F. (1996). "Upgrading travel demand forecasting capabilities-U.S. DOT travel model improvement program." *TR News*, 186.