TOWARDS AN ACTIVITY-BASED APPROACH FOR ESTIMATING TRAVEL DESTINATIONS

Shan Jiang, shanjang@mit.edu Massachusetts Institute of Technology, USA

Filipe Rodrigues, fmpr@student.dei.uc.pt Universidade de Coimbra, Portugal

Ana Alves, ana@dei.uc.pt Universidade de Coimbra, Portugal

Francisco Pereira, camara@eden.dei.uc.pt Universidade de Coimbra, Portugal

Joseph Ferreira, jf@mit.edu Massachusetts Institute of Technology, USA

ABSTRACT

Transportation demand models rely heavily on destination information. The activity-based model especially requires high resolution and disaggregated information of activity destinations. Recent developments in spatially-detailed, GIS-based data sources are making it practical to consider new methods for modeling urban activity in ways that can facilitate travel demand estimation. Massive amounts of data on land use, points of interest, public events, urban sensing, etc. are becoming available online. These data, together with modern techniques for geo-processing and data fusion, offer new possibilities for deriving activity destinations. In urban settings, such analyses can also link travel patterns with different activity patterns in ways that can be usefully incorporated into models of land use and transportation interactions. This paper develops and analyzes data fusion and estimation methods that use such data to estimate the location and size of the urban activity destinations, which are key to activity-based land use and transportation modeling. The methods are developed and illustrated using six towns in the Boston metropolitan Area, USA, as examples. Data sources include online derived points of interest from Yahoo!, proprietary business establishment data, Census Block and Census Block Group boundary data, and census employment data. This new approach for estimating activity destinations and incorporating them into travel demand can be beneficial for cities that lack current detailed business survey data for building activity-based models but wish to test the sensitivity of travel behavior to policy options and ITS implementations that are likely to alter activity patterns.

Keywords: Machine Learning, Data Fusion, Points of Interest, Activity-Based Modeling, Four-Step Modeling, Land-Use and Transportation Interaction, Destination Estimation, GIS

INTRODUCTION

Understanding personal travel patterns and modeling travel demand has been essential for planners to plan efficient urban transportation systems to fulfill mobility needs. In the past half century, this effort has been dominated by the four-step modeling (FSM) approach, which is composed of steps (1) trip generation, (2) trip distribution, (3) mode split and (4) trip assignment. The first two steps of the four-step model are to produce measures of travel demand based on the activity system (traveler characteristics and land use information), and the last two steps try to allocate the formally estimated travel demand onto the transportation network (McNally, 2008).

Recent development in spatially-detailed, GIS-based data sources are making it practical to consider new methods for modeling urban activity in ways that can facilitate travel demand estimation. Massive amounts of data on land use, points of interest (POIs), high resolution orthophotos, public events, urban sensing, etc. are becoming available online. These data, together with modern techniques for geo-processing and data fusion, offer new possibilities for deriving activity destinations. In urban settings, such analyses can also link travel patterns with different activity patterns in ways that can be usefully incorporated into models of land use and transportation interactions.

Given such a background, we propose to answer the following question:

> How can we employ data fusion methods to estimate disaggregated urban activity destinations, by using the emerging data sources (i.e., derived point-of-interest information, road networks, and land use)?

We develop and analyze data fusion methods that use such data to estimate urban activity destinations. With the use of data from six towns in the Boston metropolitan Area, USA, as examples, we develop and illustrate the methods. Data sources include employment by category at aggregated Census Block Group level, derived point-of-interest information, proprietary business establishment data, and geographical boundaries of the aggregate and disaggregate units of analysis.

This research will be beneficial for cities that lack detailed or timely survey data for building activity-based models but wish to test the sensitivity of travel behavior to policy changes, such as Intelligent Transportation Systems (ITS) implementations that are likely to alter activity patterns.

LITERATURE REVIEW: MOTIVATION

In this section, we examine the development and evolution of transportation demand models, and land use, transportation, and environmental models so as to understand the needs for developing new destination estimation methods.

Transportation Demand Models

Traditional Four-Step Models

The traditional four-step model (FSM) of travel demand and network allocation has been widely employed for decades, especially in the performance analysis of transportation systems. However, there are many limitations of the FSM approach, including the ignorance of the spatiotemporal characteristics of household travel behavior, the assumption of a fixed pattern of underlying activities, and the lack of integration with land-use forecasting models. In other words, the derived nature of the demand for transportation is not reflected well in the FSM methodology (McNally, 2008).

In such a methodological framework with imperfections, the travel-demand estimation may not reflect accurately the real world situation, especially when having to describe individual reactions to certain policy measures that may change their activity patterns. The FSM may fail to provide a robust foundation for policy analysis to tackle mobility problems that many urban areas are facing today.

Activity-Based Travel Demand Models

In contrast, the activity-based approach (ABA) has constructed a much richer framework for estimating activity patterns at the individual and household level (Ben-Akiva, Bowman, & Gopinath, 1996). The major characteristics of the ABA model are: (1) travel is derived from the demand for activities; (2) tour is used as the analysis units for travel pattern, instead of trip as in the FSM; (3) household and social structures influence activity behavior; (4) spatial temporal, mode, personal interdependencies constrain activity behavior (McNally, 2008).

Urban Simulation Models

Early Efforts at Large-Scale Urban Modeling

Beyond the aforementioned traditional transportation models (such as FSM) that were first applied to metropolitan areas in the 1950s, attempts to build spatially detailed large-scale metropolitan simulations started in the 1960s and have continued through the decades despite significant limitations and obstacles (Anas & Duann, 1985; Batty, 2003; Lee, 1973; Lowry, 1964).

Land Use, Transportation and Environmental (LUTE) Models

Recent work on large-scale urban models has focused on the construction and interconnection of theory-based subsector components that integrate economic development, environmental management and transportation elements, and utilize micro-simulations of submarkets (Ferreira, Diao, Zhu, Li, & Jiang, 2010). Examples of these dynamic systems of

LUTE models include UrbanSim (Waddell, 2002; Waddell, Wang, & Charlton, 2008), ILUTE (Salvini & Miller, 2005), DaySim (Bradley, Bowman, & Griesenbeck, 2007), etc. These urban simulation models have integrated different modules, such as a population synthesizer, land-use development, household-activity pattern, household- or firm-location-choice module, transportation-systems model, etc.

Challenges and Opportunities

As urban simulation models evolve, the demands for disaggregated data increase greatly, ranging from population data and employment data, to travel-survey data. On the one hand, employment data with detailed size, type and location are still expensive to get and not well understood. On the other hand, destinations tend to be more concentrated or clustered than residential locations. Therefore the traditional disaggregation approach, assuming uniform distribution of destinations (or employment opportunities) across space, is not plausible.

In this research, we try to utilize the emerging online public data sources (such as online point-of-interest data containing location and category information) to develop new data-fusion methods for estimating disaggregated activity destinations, which are more easily re-structured as models and conditions change.

MODEL STRUCTURE AND METHODS

The overall model structure for this study is illustrated in Figure 1. By using machine learning, data-fusion methods, we combine the employment data at the aggregated (e.g., US Census Block Group) level and the online extracted point-of-interest (POI) (locational and categorical) information to estimate employment sizes at disaggregated (e.g., US census Block) level, and compare the results with those obtained from state-of-the art proprietary business location databases.

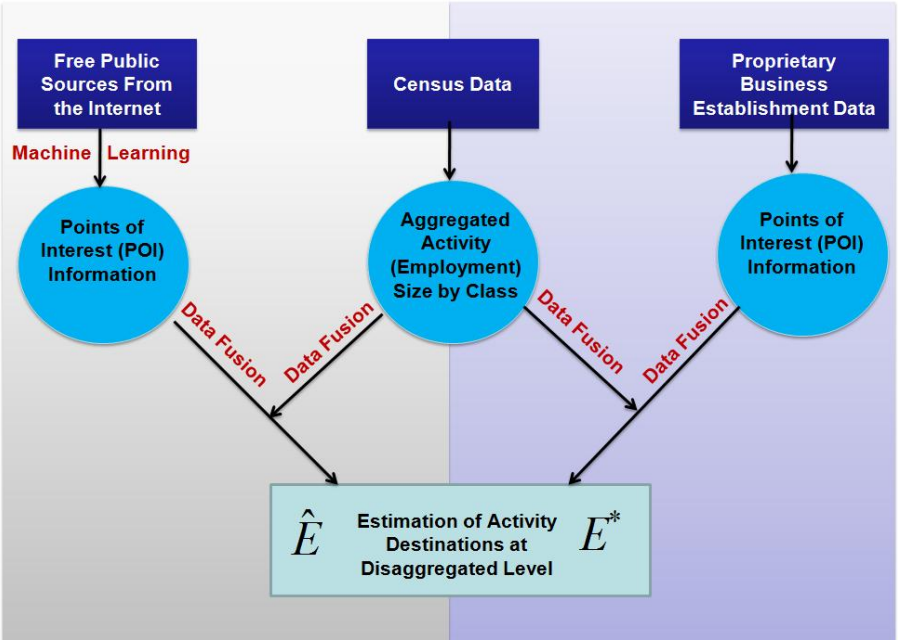

Figure 1 Illustration of model structure

Machine Learning: Capture POIs from the Internet

We first extract information of online points of interest (POIs) through public application programming interfaces (APIs) of some online user-content platforms (e.g., Yahoo! database). The information includes POI business names, coordinates, addresses, and categories.

Due to its nature, this kind of online POI database usually grows faster than proprietary POI databases, such as infoUSA business establishment database (infoUSA, 2010). However, there often exist duplicated POIs in the online user-content sources and their categorization does not follow a strict classification standard (e.g., the North American Industry Classification System--NAICS) as used in most proprietary business establishment databases. Our hypothesis is that there is considerable coherence between categories of online user-content platforms (e.g.,Yahoo!) and NAICS codes, such that a model can be trained to automatically classify incoming online extracted (e.g., Yahoo!) POIs.

Therefore we employ a matching algorithm to detect duplicates by comparing POIs according to their names, website information, and geographic distances. We make use of the JaroWinklerTF-IDF class from the SecondString project (Cohen, Ravikumar, & Fienberg, 2003) to indentify close names, ignore misspelling errors and some abbreviations.

We then use Weka (Witten & Frank, 2005), a data mining platform that provides a wide number of classification algorithms. In our experiments, we classify POIs for different NAICS levels (i.e. NAICS categories with different granularities), particularly two-, four- and six-digit NAICS codes. Two-digit codes allow us to analyze economic sectors, while six-digits specify the detailed categories of business establishments). For validation purposes we use ten-fold cross-validation (Mitchell, 1997). We also perform validation with an external test set containing POI data for a different city to understand the dependency of the model on the study area.

Data-Fusion & Maximum Likelihood Estimation (MLE)

The definition of “data fusion” varies in different research fields (Wald, 1999). One definition relevant to this research defined by Mangolini (1994) is that “data fusion is a set of methods, tools and means using data coming from various sources of different nature, in order to increase the quality (in a broad sense) of the requested information”. The Joint Directors of Laboratories (JDL) of the U.S. Department of Defence (1991) defines data fusion as a “multilevel, multifaceted process dealing with the automatic detection, association, correlation, estimation and combination of data and information from single and multiple sources”. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets (Seifert, 2006).

We use a set of POIs extracted and classified from user-content platform (e.g. Yahoo!) to disaggregate the aggregated data to a finer level, and use a proprietary source--ESRI Business Analysis package (ESRI, 2009), which contains detailed information of business

establishments provided by infoUSA in 2008 in the US--to evaluate our newly developed method.

To support LUTE modeling where travel demand is sensitive to block level travel time and distances, we would like to have destination indentified at the scale of city block level. We employ a local maximum likelihood estimation (MLE) method to disaggregate Block Group level aggregates to Block level destination estimations.

We treat employment sizes at different POIs as random variables. We assume that employment sizes of a certain category within a Block Group are independent and identically distributed (i.i.d.). Therefore, in a Block Group, the maximum likelihood estimates (MLE) of the employment sizes (of a certain category) within different Blocks are proportional to the numbers of POIs within the Blocks. In other words, the share of the estimated employment size of a Block in a Block Group is equal to the share of POIs of the Block in the Block Group. Since the XY location of POIs includes measurement error, we buffer the XY locations and treat the assignment of POIs to blocks as a random variable (as described later).

Model Evaluation

By employing the MLE method described above and using business establishment survey data (e.g., ESRI Business Analysis package), we obtain a benchmark employment size of category *c* at Block *b* in Block Group *g*, $E^{*}_{b,c,g}$, which is taken as the true value of the disaggregated employment size. By using the derived POI information (obtained from the machine learning algorithm), we obtain an ML estimate of employment size of category *c* at Block *b* in Block Group *g*, $\hat{E}_{b,c,g}$.

We then use the *mean squared error* (MSE), a commonly used measurement, to quantify the difference between an estimator and the true value of the quantity being estimated. In order to compare our method with the traditional disaggregation approach (assuming uniform distribution of employment opportunities), we use the ratio of MSEs of our MLE method and the traditional uniform disaggregation method, the relative mean squared error (RMSE), to evaluate the goodness of fit of our model, which will be discussed later in detail in the case study section.

STUDY AREA AND DATA

Study Area

The main purpose of this research is to develop and test a new method for estimating destination data at a disaggregated level for metropolitan areas that are keen to develop activity-based transportation models or agent-based urban simulation models. To do this, we need to start with a city where all the data (such as GIS, and business establishment data, etc.) required to develop, calibrate, and validate the proposed new model are available. On

the other hand, in order to test the robustness of the method, we also want to include different cities.

Based on these rationales, we selected 6 towns located within the first ring road (Route 128) in the Boston metropolitan Area (Figure 2). This area stretches from the core of the Boston metro area to the edge of the first major circumferential interstate highway in the metropolitan area. Table 1 describes the population, area, and employment in these 6 towns.

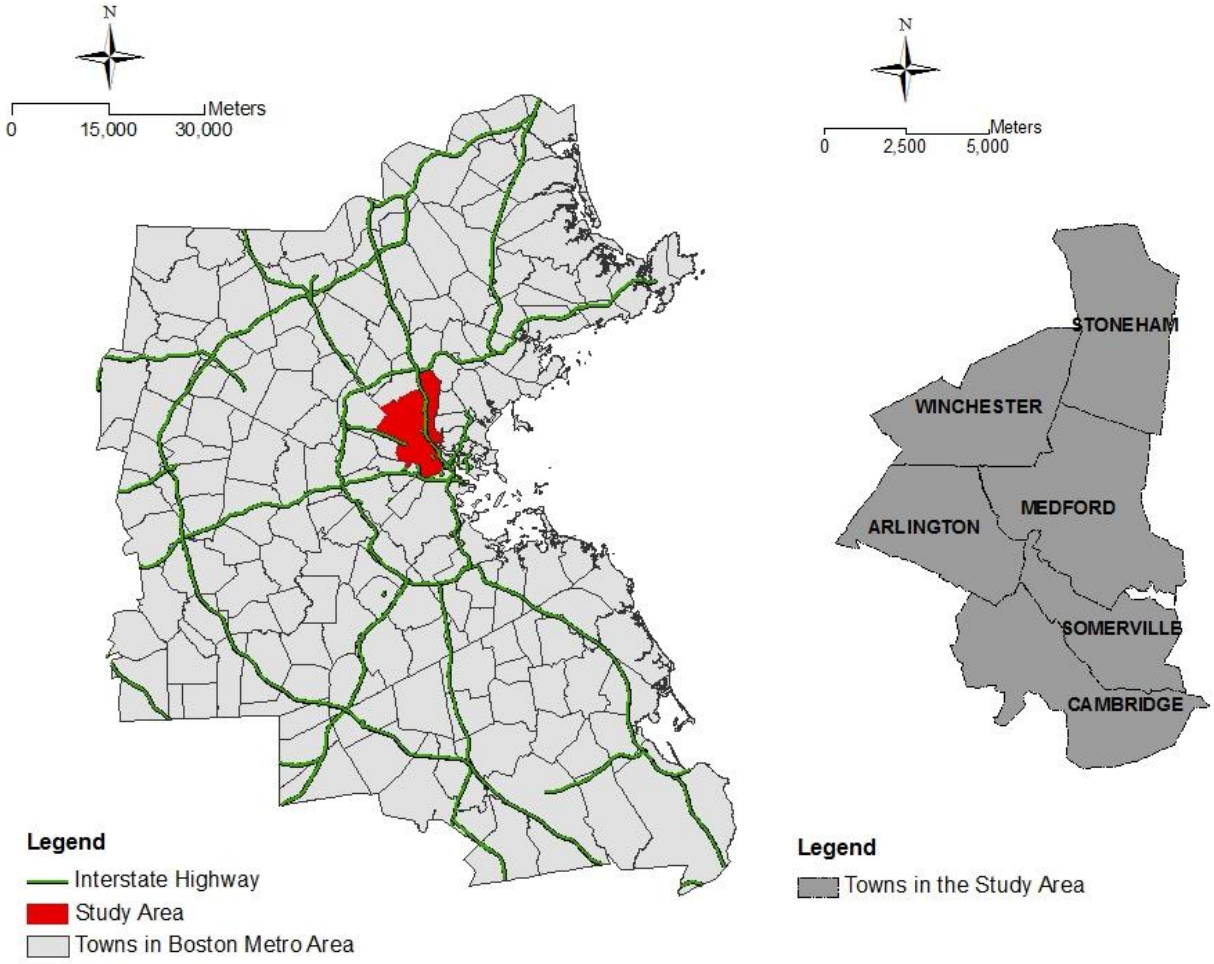

Figure 2 Boston Metropolitan Areas and 6 Selected Towns in the Study Area

Table 1 Population and employment size and density of the 6 selected towns in Boston Metro Area

Town Name	Population, 2000	Employment, 2000	Area (sq km)	Population Density (residents per sq km)	Employment Density (workers per sq km)
Arlington	42,389	132,178	14.27	2,969	9,259
Cambridge	101,355	1,803,758	18.47	5,489	97,683
Medford	55,765	298,380	21.88	2,549	13,640
Somerville	77,478	359,607	10.65	7,275	33,766
Stoneham	22,219	163,070	17.38	1,278	9,383
Winchester	20,810	224,146	16.30	1,277	13,753

Data Source: U.S. Census 2000 and MassGIS

Due to the substantial efforts of MassGIS (the Commonwealth's Office of Geographic and Environmental Information), there are ample GIS data for the Boston metro area available for public use. In addition, we also utilized the latest (2008) proprietary business establishment data for Boston metro area (from ESRI Business Analysis package, containing 2008 infoUSA data), which is crucial to the model validation. The development of this new method will help derive disaggregated destination estimations (measured by disaggregated employment size by category), which will facilitate the urban modeling efforts undertaken by local agencies.

Data

Data sources for this study include the following:

- Point-of-interest (POI) information derived from sources on the Internet (e.g., Yahoo!, and Dun & Bradstreet)
- Employment by category data at the Block Group level obtained from the 2000 Census Transportation Planning Products (CTPP) database
- GIS data for the boundaries of Towns, Block Groups, and Blocks downloaded from MassGIS public online data sources
- 2008 InfoUSA business establishment data included in the ESRI Business Analyst data package, which is used for model evaluation

Points of Interest (POIs)

Our data consists of a large set of POIs extracted from Yahoo! through their public API and another training set originally developed by Dun & Bradstreet (Dun & Bradstreet, 2010), a consultancy company that specializes in commercial information and insight for businesses. In the former case, the database is essentially built from user contributions; in the latter, the data acquisition process is semi-automatic and involves integration of official and corporate databases, statistical analysis and manual evaluation (Dun & Bradstreet, 2010). In both cases, information of a POI includes a name, a XY location, and a set of categories. The POIs from D&B have (2007 version) NAICS codes, but the ones from Yahoo! do not. Each POI from Yahoo! is assigned, on average, roughly two categories from the Yahoo! taxonomy of business types.

We extract 64133 POIs from Yahoo! for the Boston metropolitan area within the first ring road (Route 128) of the metro area, and 29402 from the database developed by D&B for the same area. This D&B dataset was used for training our algorithm to match NAICS categories. We estimate that the Yahoo's category taxonomy has more than 1300 distinct categories distributed along a 3-level hierarchy. After employing the matching algorithm mentioned in the machine learning section above, we build a database where a point of interest (POI) contains a set of categories and a NAICS classification. From the database developed by D&B, our data covers 514 distinct six-digit NAICS codes. However, the 2007 NAICS taxonomy has a total of 1175 six-level categories, meaning that our sample data only covers the most common NAICS codes.

An analysis on the coherence between NAICS and Yahoo! shows that only 80.2% of the POIs have consistent corresponding NAICS with the most common one for the given set of categories. For the two and four-digit NAICS, the matching consistencies are 87.1% and 83.4%, respectively. Therefore, by having the same set of Yahoo! categories mapping to different NAICS codes in different occasions, we do not expect to obtain a perfect model that classifies all cases correctly. However, the purpose of this research is not to focus on the exact matching between the online Yahoo! POIs and the standard classification provided by the proprietary business establishment data sources. Our focus is to test if the dataset generated from online POIs (even though it's not complete) can be useful in developing reasonably good disaggregated destination estimations.

At the two-digit NAICS code level, there are 20 categories of sectors. In this paper, we focus on the POIs in the retail sector (two-digit NAICS code= 44 or 45) as a demonstration due to space limitation. Figure 3 displays the retail POIs obtained from Yahoo! (left) and infoUSA (right). Table 2 and Table 3 summarize the number of POIs by 3-digit NAICS category and by town. We can see that the online extracted POIs only identify 50% of the total POIs (listed in certain categories by the proprietary business establishment source), and this ratio varies across categories and towns.

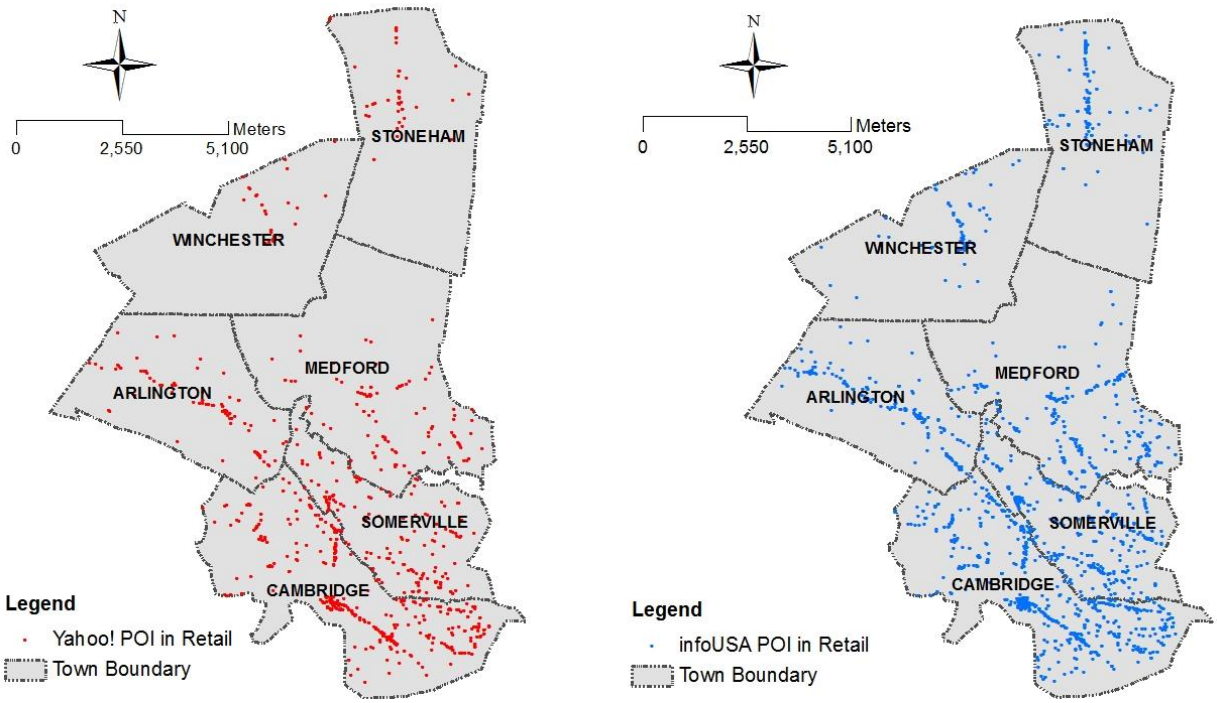

Figure 3 Distribution of Retail POIs from Yahoo! (left) and infoUSA (right) in the Study Area

Table 2 Statistics of retail POIs in the study area from Yahoo! and infoUSA by NAICS 3-Digit classification

NAICS 3-Digit Code	NAICS Description	infoUSA Count	Yahoo! count	Yahoo! to infoUSA ratio
441	Motor Vehicle and Parts Dealers	96	79	82.3%
442	Furniture and Home Furnishings Stores	104	57	54.8%
443	Electronics and Appliance Stores	251	224	89.2%
444	Bldg. Material+ Garden Equip.+ Supplies Dealers	104	90	86.5%
445	Food and Beverage Stores	268	n.a.	n.a.
446	Health and Personal Care Stores	130	44	33.8%
447	Gasoline Stations	79	n.a.	n.a.
448	Clothing and Clothing Accessories Stores	267	157	58.8%
451	Sporting Goods, Hobby, Book, and Music Stores	175	108	61.7%
452	General Merchandise Stores	61	n.a.	n.a.
453	Miscellaneous Store Retailers	301	171	56.8%
454	Non-store Retailers	17	4	23.5%
Total		1,853	934	50.4%

Note: n.a.= not available

Table 3 Statistics of retail POIs in the study area from Yahoo! and infoUSA by Town

Town Name	infoUSA Count	Yahoo Count	Yahoo to infoUSA ratio
Arlington	174	104	59.8%
Cambridge	830	438	52.8%
Medford	301	131	43.5%
Somerville	340	165	48.5%
Stoneham	113	55	48.7%
Winchester	93	35	37.6%

Aggregated Retail Employment Data

The choice of spatial analysis units at the aggregated level (i.e., Transportation Analysis Zone, or Census Tract, or Census Block Group, etc.) depends on the availability of data. For example, the employment-by-category data for our study area (6 selected towns in the Boston metropolitan area) are available at both the Census Block Group level and the Transportation Analysis Zone (TAZ) level in the Census Transportation Planning Products (CTPP) database Part II. As the resolution at the Block Group (BG) level is higher than that at the TAZ level (for the 2000 census), we disaggregate the census employment data from Block Group level to Block level by using the extracted online POIs.

The CTPP database distinguishes 14 major categories of employment (e.g., agriculture, construction, manufacturing, wholesale, transportation, information industry, finance industry, professional services, educational industry, recreation and food service industry, etc.), of which retail is one. These 14 categories correspond to the combined categories of the NAICS two-digit system. For example, the retail category in the CTPP database correspond to a NAICS two-digit code 44 or 45.

Figure 4 shows the Block Group level retail employment density in the 6 selected towns in the Boston metro area. At this stage, employment densities for different Blocks within the same Block Group are the same (see Figure 4), since we have not yet used POI locations to differentiate the Blocks within a Block Group. Table 4 describes the numbers of Block Groups and Blocks in the 6 selected towns.

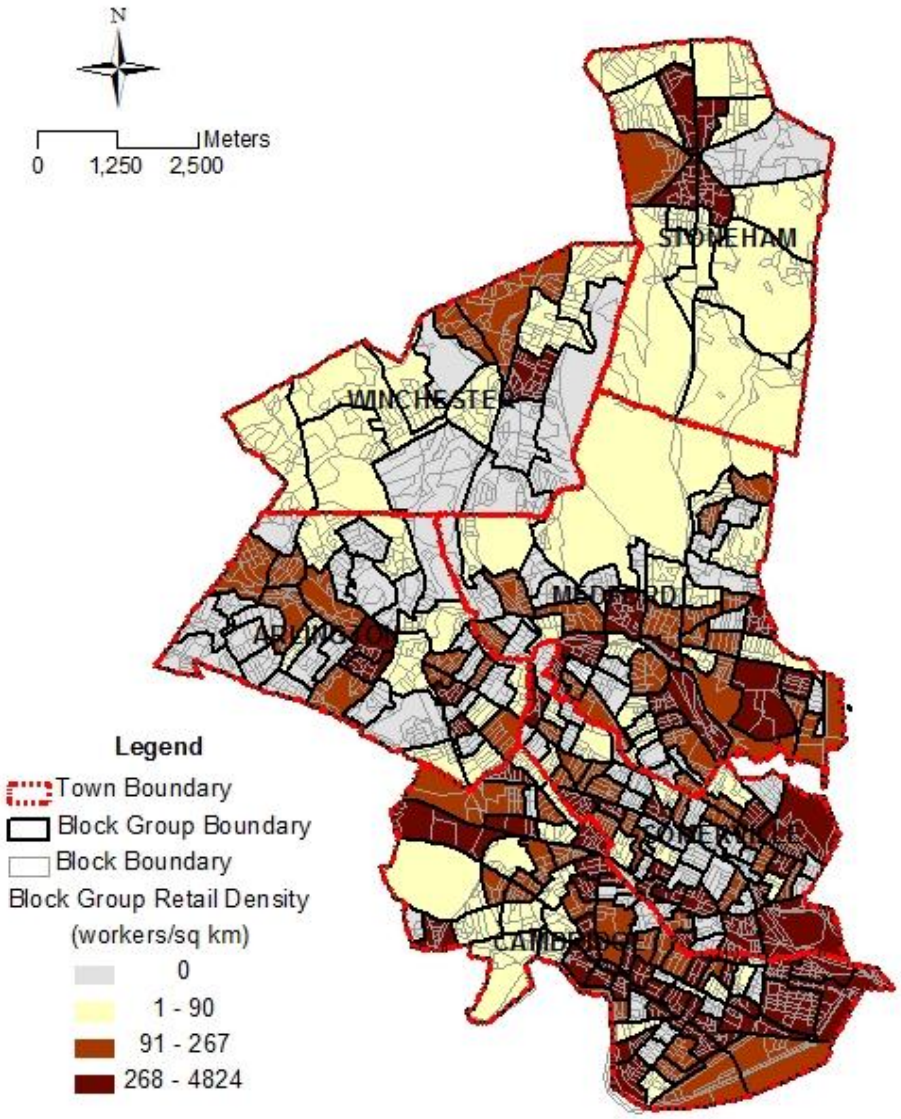

Figure 4 Aggregated employment densities at the Block Group level.

Table 4 Number of Block Groups and Blocks in the 6 Towns

Town Name	# of Block Groups	# of Blocks	Average # of Blocks in a BG
Arlington	44	651	15
Cambridge	81	886	11
Medford	57	736	13
Somerville	67	693	10
Stoneham	16	300	19
Winchester	15	377	25

Data Source: U.S. Census 2000 and MassGIS

Data Validity and Reliability

One of the data validity threats comes from the geographical information of the points of interest (POIs). First, the proposed new method depends heavily on the geo-coded locations of POIs. However, in current GIS systems, points with (X,Y) coordinates are usually geo-coded along central lines of roads, which may offset some distance from boundaries of selected geographic analysis units (such as Block Groups). The same POI in different database sources may also have different geo-locations, due to geo-coding errors. Thus systematic measurement errors may exist within the same source, and across different sources. Therefore, incorporating methods that can reduce such kind of errors is very important to the reliability of this study.

In order to address the problem of potential geo-coding errors, we create a buffer area with a 25-meter radius for each POI, and use the area share of each POI buffer in a block as the probability that each POI may fall into that block. The 25-meter size is determined by the relative road width and block size—we want the buffer size to be large enough to cover both sides of the road, but not too large to cover the entire block at each side.

MODEL ESTIMATION AND RESULTS

Figure 5 demonstrates the modeling processes of the estimation of employment size and density by category at the disaggregated (e.g., Block) level. These processes follow the same general methods as described earlier in the model structure and methods section.

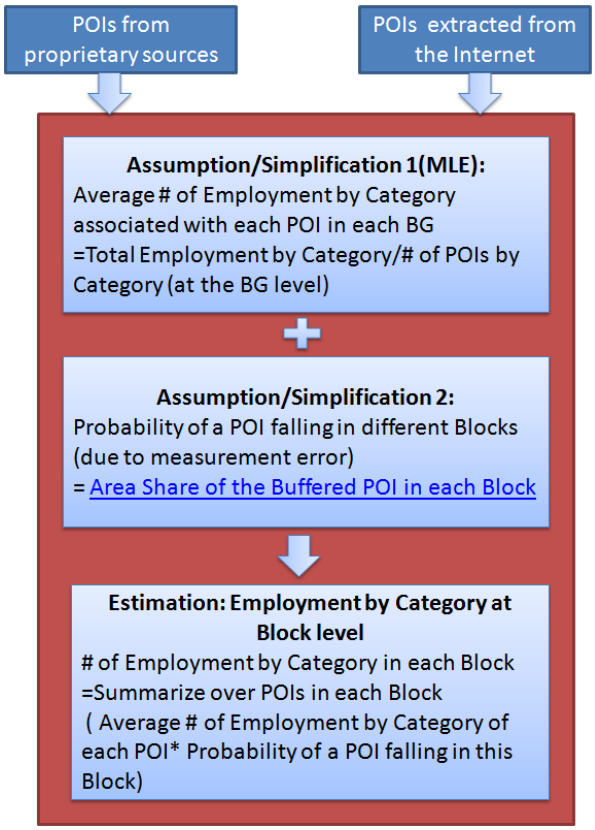

Figure 5 Demonstration of the model estimation processes.

Figure 6 and Figure 7 show the estimation results of the disaggregated retail employment density at Block level in the 6 towns of our study area, by using the two different sources of POI data (infoUSA, and Yahoo!). By comparing the estimation results, we find that the disaggregated employment estimations by using the Yahoo! POIs and those obtained from the proprietary source (infoUSA 2008) are very close.

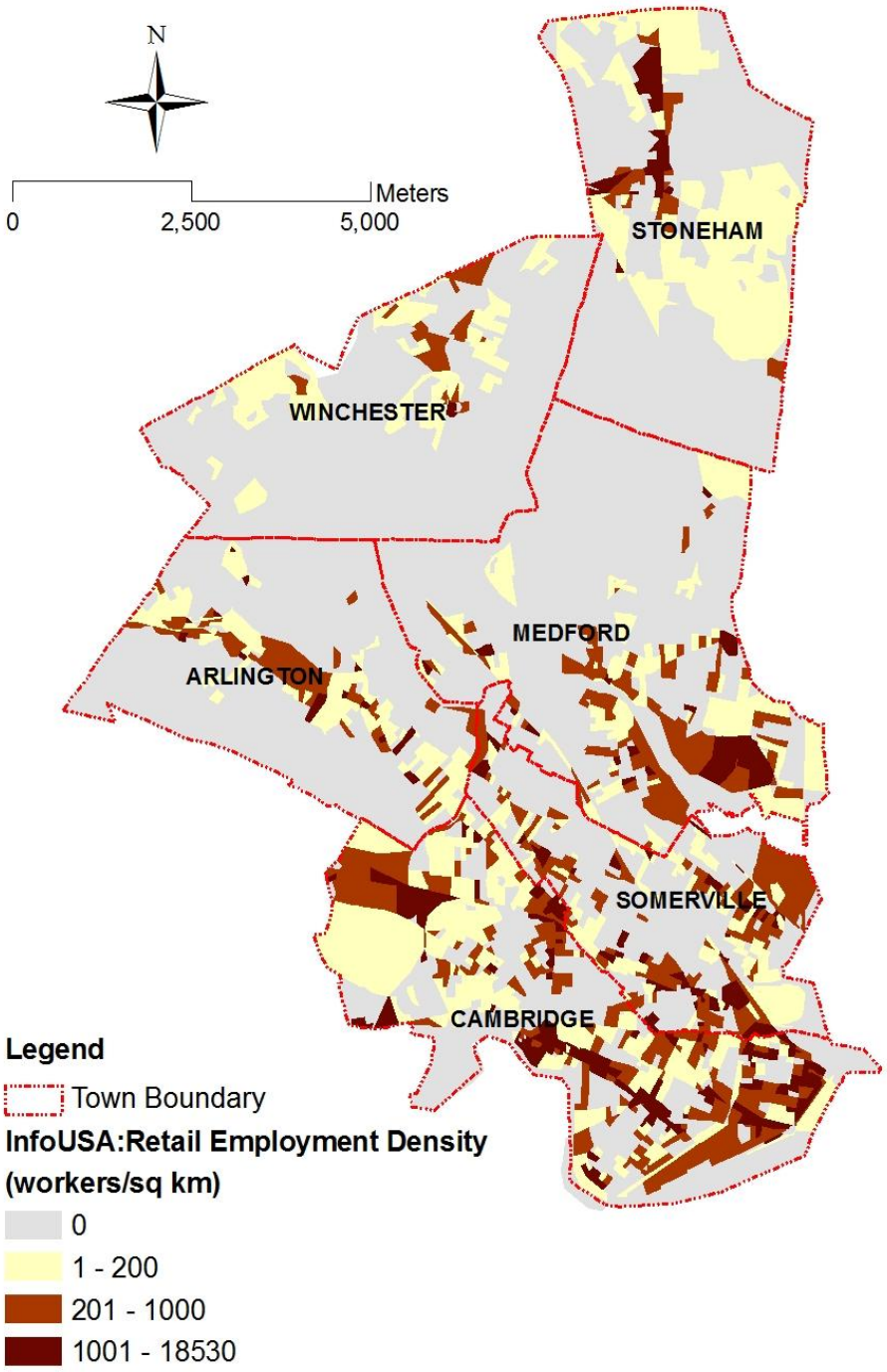

Figure 6 Estimated disaggregated retail employment density at Block level by using infoUSA POIs

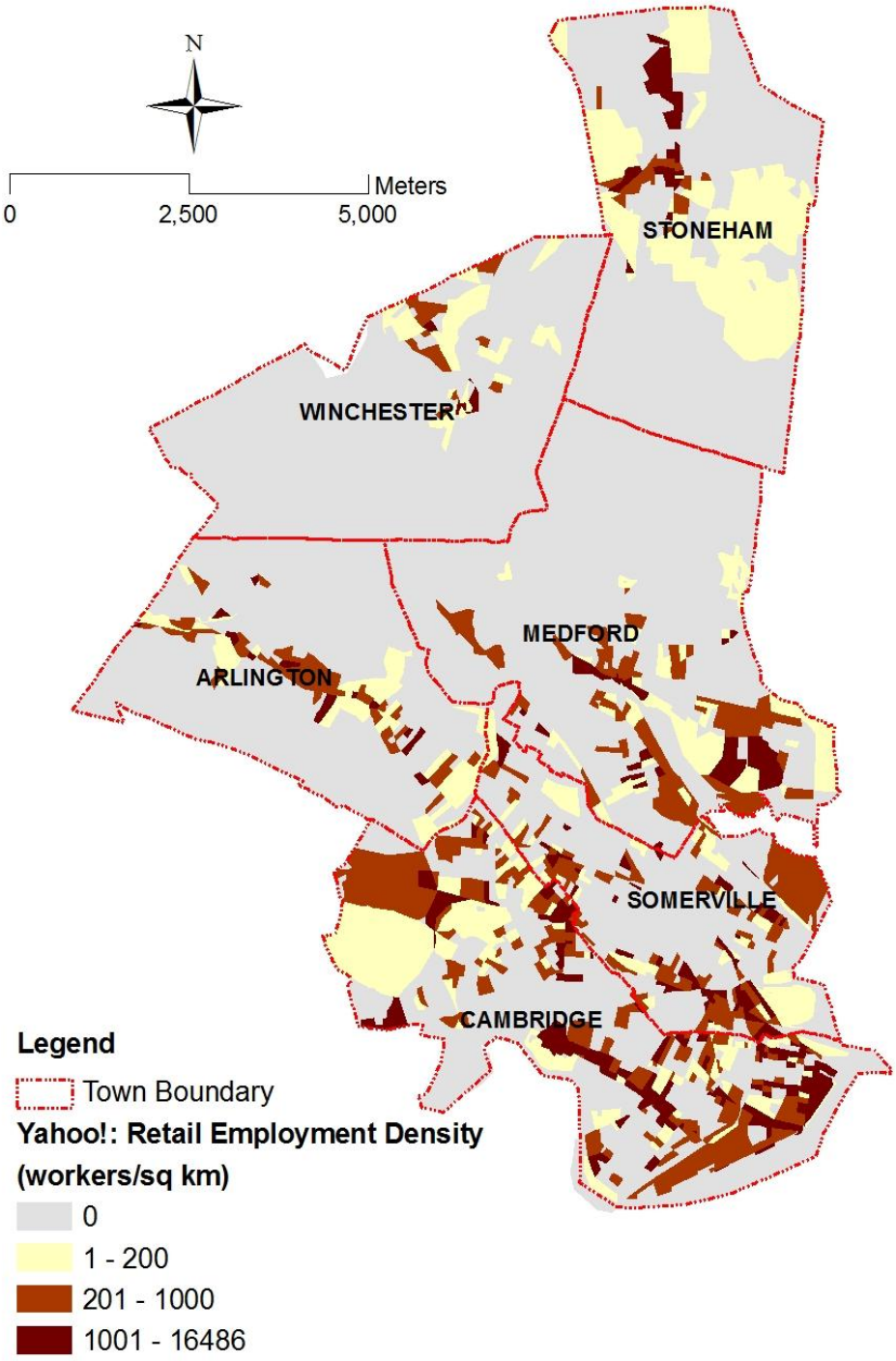

Figure 7 Estimated disaggregated retail employment density at Block level by using Yahoo! POIs

We are also interested in quantitative evaluations of the estimation results. We have introduced, in the model structure and methods section, the commonly used statistical index, the *relative mean squared error* (RMSE). Here, we give the rigorous mathematical form of RMSE (see Equations 1 and 2) to evaluate the goodness of fit of the model.

$$\overline{E}_{b,c,g} = \frac{w_{b,g} E^*_{c,g}}{\sum_q w_{q,g}} \qquad \text{[Equation 1]}$$

$$RMSE(\hat{E}, E^*) = \frac{\sum_{b,c,g} \left(\hat{E}_{b,c,g} - E^*_{b,c,g}\right)^2}{\sum_{b,c,g} \left(\overline{E}_{b,c,g} - E^*_{b,c,g}\right)^2} \qquad \text{[Equation 2]}$$

In Equation 1, $w_{b,g}$ is the area of Block *b* in Block Group *g*; $E^*_{c,g}$ is the aggregated true value of employment size of category *c* in Block Group *g*; and $\overline{E}_{b,c,g}$ is the estimated employment size at Block *b* of category *c*, using the traditional disaggregation approach, which assumes that the employment is uniformly distributed across Blocks in each Block Group *g*.

In Equation 2, $E^*_{b,c,g}$ is the benchmark employment size of category *c* at Block *b* in Block Group *g*, viewed as the true value of the disaggregated employment size derived from the proprietary business establishment data source; $\hat{E}_{b,c,g}$ is the maximum likelihood estimate (MLE) of employment size of category *c* at Block *b* in Block Group *g*, employing the extracted online Yahoo! POIs.

The *relative mean squared error* (RMSE) is the ratio of the mean squared error (MSE) using the data-fusion method to the MSE using the traditional Block Group average estimation method. If the RMSE is less than 1, it means that the data-fusion method using the derived POIs improves the destination disaggregation estimates; the smaller the RMSE, the greater improvements the data-fusion method makes. If the RMSE is close to 0, it means that the data-fusion method using the extracted online POIs gives very precise estimates. However, if the RMSE is greater than 1, it means that the derived POIs do not well reflect the distribution of the population POIs (as listed in the proprietary business establishment database).

As described in the POI data description section, we notice that our extracted online POIs do not match perfectly with the proprietary business establishment data. However, we conjecture that, on average, the estimations of disaggregated employment at block level will be improved compared to the traditional uniform disaggregation approach—as these POIs, to some degree, represent the distribution of activity destinations across space and reflect their heterogeneous nature.

Employing Equation 2, the disaggregated employment estimation at the Block level using Yahoo! POI gives

RMSE = 0.407.

The RMSE is significantly smaller than 1, which means that using the extracted Yahoo! online POIs to estimate the disaggregated employment sizes at the Block level has reduced the mean squared error by around 60% compared to the traditional uniform disaggregation approach.

We also conjecture that the improvement in the estimation of disaggregated employment in large blocks is more significant than that in small blocks, compared to the traditional uniform disaggregation approach. The underlying reasons are the following:

(1) the impacts of POI geo-coded errors in blocks with large areas are relatively smaller than those in blocks with small areas;
(2) the relative gaps between the extracted online Yahoo! POIs and those obtained from the proprietary infoUSA database in blocks with small areas are larger than those in blocks with large areas, as blocks with larger areas usually contain more POIs, and the geo-coded errors matter less (since street width is a small fraction of block size).

We sort the 3633 Blocks (with complete data within our study area) by their areas, and divide them into two groups—one consisting of 1817 blocks with smaller areas, and the other consisting of 1816 blocks with larger areas. We compute the RMSE for each group, and the RMSE for the group with smaller block sizes is 0.570, and RMSE for the group with larger block sizes is 0.394. These results accord with our previous conjecture.

CONCLUSIONS

According to our case study, it has shown that by using the data-fusion methods developed in this paper (combining POIs and aggregated data), we can derive more accurate estimations of activity destinations (such as retail destination measured by retail employment size) at disaggregated level (e.g., U.S. Census Block level), compared to the traditional uniform disaggregation approach which assumes uniform distribution of destinations across Block Groups. The data-fusion methods can be applied by using both proprietary business establishment data and online point-of-interest (POI) data.

In general, several issues of data validity and reliability exist for the extracted online point-of-interest (POI) data; however, these data are still very useful in estimating more accurate disaggregated destinations. First, as for the POI information extracted from the Internet, the coverage and accuracy of this information depends heavily on (1) the completeness of online public sources, and (2) the consistency of public categories. For most urban areas in the U.S., where information technology (i.e., the Internet) has been widely used to provide and acquire information, the POI information can be widely accessible, but potential gaps may still exist between the total business establishments and available information online. These gaps will be reduced as more cities improve their information technology infrastructure, and the online user-content platforms for publishing POI information apply rigorous and standardized categorization guidelines. On the other hand, combining different online sources may help to reduce these gaps, but may also introduce problems of redundancy. This is also one of the potential issues that our machine learning method has tried to deal with.

By using machine learning algorithms and online point-of-interest (POI) information, we can estimate activity destinations with richer information. For cities without resources to purchase or update business establishment data (such as infoUSA data), the machine learning method and data-fusion method provide an alternative possibility for developing timely disaggregated travel destination estimations, which are essential for activity-based travel demand models and much better than could be done using Block Group level data alone.

For example, we have applied similar disaggregation approach (using extracted online POIs and aggregated employment data) in Lisbon, Portugal, where an integrated Land Use, Transportation and Environment (LUTE) model is being built by the transportation focus areas of the MIT-Portugal program (MPP). By using the proposed methods to derive disaggregated destination estimation, we plan to test components in the integrated LUTE model for Lisbon that account for activity-based transportation demand and support household level micro-simulation, etc.

Disaggregated destination information is very important to improve travel demand modelling (including the traditional four-step model, and the newly developed activity-based model), as travel demand is very sensitive to micro-level changes in travel time and travel distances. Since destinations tend to be more clustered and concentrated than residential locations, the location and categorization information of points of interest (POIs) is very useful for us to understand the destination characteristics and the derived travel demand at the micro-level. The data-fusion methods developed in this research provide us with new possibilities to study the micro-level travel behavior and travel demand, and open a new window for cities with limited resources that wish to develop policy-sensitive transportation demand models at the disaggregated level.

ACKNOWLEDMENTS

We acknowledge the MIT Portugal Program for supporting this work on information infrastructure and geo-processing for transportation and land use modeling. In addition, we appreciate the computing facilities and IT assistance at both MIT's Computer Resource Network (CRON) and the Universidade de Coimbra.

REFERENCES

[JDL] Data Fusion Subpanel of the Joint Directors of Laboratories Technical Panel for C3. (1991). Data fusion lexicon. In U.S. Department of Defence (Ed.).

Anas, A., & Duann, L. S. (1985). Dynamic forecasting of travel demand, residential location and land development. *Papers in Regional Science, 56*(1), 37-58.

Batty, M. (2003). New Developments in Urban Modeling: Simulation, Representation, and Visualization. In *Integrated Land Use and Environmental Models: A Survey of Current Applications and Research* (pp. 13-46). New York: Springer.

Ben-Akiva, M., Bowman, J. L., & Gopinath, D. (1996). Travel Demand Model System for the Information Era. *Transportation Research Part A: Policy and Practice, 23*(3), 241-266.

Bradley, M. A., Bowman, J. L., & Griesenbeck, B. (2007). *Development and application of the SACSIM activity-based model system.* Paper presented at the 11th World Conference on Transport Research, Berkeley, California, USA.

Cohen, W., Ravikumar, P., & Fienberg, S. (2003). *A comparison of string distance metrics for name-matching tasks.* Paper presented at the Proceedings of the IJCAI-2003 Workshop on Information Integration on theWeb (IIWeb-03), Acapulco, Mexico.

Dun & Bradstreet. (2010). D & B website. from http://www.dnb.com/

ESRI. (2009). ArcGIS Business Analyst Package.

Ferreira, J., Diao, M., Zhu, Y., Li, W., & Jiang, S. (2010). Information Infrastructure for Resesarch Collaboration in Land Use, Transportation, and Environmental Planning, *The Transportation Research Board (TRB) 89th Annual Meeting*. Washington, D.C.

infoUSA. (2010). U.S. Business Lists. from http://leads.infousa.com/USBusinesses.aspx

Lee, D. B. (1973). A Requiem for Large Scale Modeling. *Journal of the American Institute of Planners, 39*(3), 163-178.

Lowry, I. S. (1964). *A Model of Metropolis*. Santa Monica, Cal.: The RAND Corporation

Mangolini, M. (1994). *Apport de la fusion d'images satellitaires multicapteurs au niveau pixel en t´el´ed´etection et photo-interpr´etation.* MSc Thesis, University of Nice–Sophia, Antipolis, France.

McNally, M. G. (2008). The Four Step Model. In D. A. Hensher & K. J. Button (Eds.), *Handbook of Transportation Modeling*. Amsterdam; London: Elsevier.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Salvini, P. A., & Miller, E. J. (2005). ILUTE: An Operational Prototype of a Comprehensive Microsimulation Model of Urban Systems. *Networks and Spatial Economics, 5*, 217-234.

Seifert, J. W. (2006). Data Mining: An Overview. In D. D. Pegarkov (Ed.), *National security issues* (pp. 201-218). New York: Nova Science Publishers.

Waddell, P. (2002). UrbanSim: Modeling Urban Development for Land Use, Transportation and Environmental Planning. *Preprint of an article in the Journal of the American Planning Association., 68*(3), 297-314.

Waddell, P., Wang, L., & Charlton, B. (2008). Integration of Parcel-Level Land Use Model and Activity-Based Travel Model, *TRB 87th Annual Meeting Compendium of Papers DVD*. Washington D.C.: TRB.

Wald, L. (1999). Some terms of reference in data fusion. *Geoscience and Remote Sensing, IEEE Transactions on, 37*(3), 1190-1193.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques* (2nd ed.): Morgan Kaufmann.