Computers, Environment and Urban Systems 53 (2015) 36–46

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Mining point-of-interest data from social networks for urban land use classification and disaggregation

Shan Jiang [a,*], Ana Alves [b], Filipe Rodrigues [c], Joseph Ferreira Jr. [d], Francisco C. Pereira [e]

[a] Department of Urban Studies and Planning, Massachusetts Institute of Technology, 77 Massachusetts Avenue Room 9-536, Cambridge, MA 02139, USA
[b] Center of Informatics and Systems, Universidade de Coimbra & Polytechnic Institute of Coimbra, Coimbra, Portugal
[c] Center of Informatics and Systems, Universidade de Coimbra, Coimbra, Portugal
[d] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, USA
[e] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

ARTICLE INFO

Article history:
Available online 12 January 2015

Keywords:
Information extraction
Machine learning
Points of interest
Land use
Volunteered geographic information

ABSTRACT

Over the last few years, much online volunteered geographic information (VGI) has emerged and has been increasingly analyzed to understand places and cities, as well as human mobility and activity. However, there are concerns about the quality and usability of such VGI. In this study, we demonstrate a complete process that comprises the collection, unification, classification and validation of a type of VGI—online point-of-interest (POI) data—and develop methods to utilize such POI data to estimate disaggregated land use (i.e., employment size by category) at a very high spatial resolution (census block level) using part of the Boston metropolitan area as an example. With recent advances in activity-based land use, transportation, and environment (LUTE) models, such disaggregated land use data become important to allow LUTE models to analyze and simulate a person's choices of work location and activity destinations and to understand policy impacts on future cities. These data can also be used as alternatives to explore economic activities at the local level, especially as government-published census-based disaggregated employment data have become less available in the recent decade. Our new approach provides opportunities for cities to estimate land use at high resolution with low cost by utilizing VGI while ensuring its quality with a certain accuracy threshold. The automatic classification of POI can also be utilized for other types of analyses on cities.

1. Introduction

Increasing amounts of data on points of interest (POIs), public events, and urban sensing are becoming available online. Spatially detailed and volunteered geographic information (VGI) together with modern techniques for geo-processing offer new possibilities for deriving disaggregated land use data that represent activities in cities. In urban settings, such analyses can link travel with different activity patterns in ways that can be usefully incorporated into models of land use and transportation interactions. As urban simulation evolves into more sophisticated activity-based land use, transportation, and environment (LUTE) models, the demand for spatially high-resolution data increases greatly. For example, with respect to work location choice, the traditional disaggregation approach that assumes uniform distribution of employment across space in the area of analysis, such as transportation analysis zones (TAZ), is no longer satisfactory. To obtain disaggregated employment data to represent detailed land use and then model transportation demand and its environmental impacts, cities have been collecting business establishment data from proprietary data sources or unemployment insurance databases (Wang, Waddell, & Outwater, 2011), which are often expensive and/or have use restrictions. Given this background, we propose to answer the following question: *How can we utilize publicly available emerging VGI sources and traditional aggregate census data to estimate disaggregated urban land use (or employment size by category)?*

In this study, we develop and apply methods that efficiently transform VGI into standardized information that can be utilized in urban planning and particularly illustrate the concept by estimating urban land use (i.e., employment size by category). We propose the use of web mining and machine learning techniques to automatically collect and classify POIs from different sources to a standard taxonomy such as the North American Industry Classification System (NAICS) (2012) used in the U.S., Canada and Mexico, which is essential for proper analysis of the POI

* Corresponding author. Tel.: +1 857 654 5066; fax: +1 617 253 3625.
E-mail addresses: shanjang@mit.edu (S. Jiang), ana@dei.uc.pt, aalves@isec.pt (A. Alves), fmpr@dei.uc.pt (F. Rodrigues), jf@mit.edu (J. Ferreira Jr.), camara@mit.edu (F.C. Pereira).

http://dx.doi.org/10.1016/j.compenvurbsys.2014.12.001
0198-9715/

data, especially when these POIs are collected from different sources. After comparing several classification methods, we apply the results to estimate employment sizes by category at a disaggregated level. With six towns in the Boston metropolitan area as an example, we develop and illustrate the methods. Data sources for this study include employment by category at the aggregate census block group level, volunteered POI information, and geographic boundaries of both the aggregate and disaggregated area of analysis. We also employ two sets of proprietary business establishment data—the first for supervised machine learning training and the second for developing a benchmark model to compare results.

The basic unit of information collected from social networks in our study is a point-of-interest, which is a specific point location that a considerable group of people find useful or interesting. The POIs are scattered across a myriad of different websites, systems and devices, complicating the development of an exhaustive database of such information. There are currently hundreds, if not thousands, of voluntarily generated POI directories on the Web, such as Yahoo! Local,[1] Google Places[2] and Facebook Places,[3] each of which uses its own taxonomy of categories or tags. To take full advantage of these resources, categories must be unified via a common taxonomy, thus maximizing the quantity and heterogeneity of information available. The type of information provided by each source can vary widely, but there exists a common set of fields generally available for each POI, including name, address (and/or GPS position), categories or tags, and optional official website. A clear distinction can be made between local business directories and platforms that are based on social networks. In the first group, the owner itself usually creates the POI and its details. The information provided is usually more accurate because the business establishment wants to be easily found by potential customers; thus, the chosen categories are more precise. In the second group, each individual is free to create a third-party reference to a geo-referenced company or service. The information entered is not validated by any authority, and, in cases with open taxonomies, the category set can be random. However, there are advantages with this group—the database is often more comprehensive in terms of number and types of POIs. Additionally, this group can have a wider spatial coverage, with the same database consisting of data from several cities or even countries; the data reflect the perspective of the user, which is relevant for leisure trip generation, for example.

In this study, we give special attention to Yahoo! data, which were essentially built from user-volunteered contributions, for two major reasons: (1) the data are comprehensive for our study area, and (2) the legal terms of Yahoo! service allow us to implement the study. The latter point is particularly relevant, as there are many such data sources on the Web with restrictive rules (e.g., massive data usage not permitted even for research). We also use proprietary business databases available for this research first for supervised machine learning and then for ground truth, including the Dun & Bradstreet (D&B)[4] and infoUSA[5] commercial databases created by consultancy companies that specialize in commercial information and insights for businesses.[6] The proprietary business establishment data acquisition process was usually semi-automatic and involved the integration of official and corporate databases, statistical analysis and manual evaluation.[7]

[1] http://local.yahoo.com (Last visited in December 2012).
[2] http://www.google.com/places/ (Last visited in December 2012).
[3] https://www.facebook.com/about/location (Last visited in December 2012).
[4] http://www.dnb.com/ (Last visited: December, 2012).
[5] http://www.infousa.com/ (Last visited: December, 2012).
[6] We obtained these two proprietary databases through university research funding.
[7] According to companies who sell these proprietary databases – http://www.dnb.com/lc/sales-marketing-education/data-quality.html and http://www.infousa.com/data-quality/.

The rest of this paper is organized as follows. Section 2 presents relevant literature and previous studies. Section 3 gives an overview of the research framework and study area. Section 4 discusses the machine learning process and results of automatically classifying online POIs. Section 5 describes the method of employing POIs and aggregate employment data to disaggregate land use into a level with higher spatial resolution and compares estimation results from VGI-based POIs with those from an independent proprietary POI source for ground truth. Section 6 concludes the paper.

2. Related work

2.1. Volunteered geographic information from social networks

The potential of location-based (LB) social networks (like Gowalla, Foursquare, and Facebook Places) has already been demonstrated in recent studies and is increasingly exploited as the dimensions of such services grow (e.g., Cranshaw, Hong, & Sadeh, 2012; Long, Jin, & Joshi, 2012). Cheng, Caverlee, Lee, and Sui (2011) provide an assessment of human mobility patterns by analyzing the spatial, temporal, social, and textual aspects associated with the hundreds of millions of user-driven footprints (i.e., "check-ins") that people leave with these services. Noulas, Scellato, Mascolo, and Pontil (2011) provide a similar study but also analyze activity and place transitions. Both of these studies are interesting and motivate a further exploitation of this type of LB service. For example, Berjani and Strufe (2011) exploit Gowalla data to develop a recommender system for places in LB Online Social Network (OSN) services based on the check-ins of the entire user base. Beyond using VGI from social networks, other studies also correlate it with sensor data such as mobile phone activity (Toole, Ulm, González, & Bauer, 2012) or taxi GPS traces (Yuan, Zheng, and Xie, 2012) to detect land use or find functional regions in the urban area. Along with the growing interest in analyzing VGI for understanding human mobility and urban dynamics, concerns about existing limited efforts in documenting and obtaining VGI with high quality and validity have been raised by researchers in geography and other fields (Elwood, Goodchild, & Sui, 2012).

2.2. Machine learning for POI classification

The applications of machine learning algorithms in classification tasks are vast and cover such diverse fields as speech recognition, economic forecasting, environmental engineering, and road traffic prediction. In urban planning, automatic approaches to classifying land use have been developed using different techniques (e.g., Griffin, Huang, & Halverson, 2006; Santos & Moreira, 2006; Currid & Connolly 2008). To use POI data to analyze cities and land use, POI classification is essential. For example, using a large commercial POI database, Santos and Moreira (2006) create and classify location contexts using decision trees, identifying clusters via a density-based clustering algorithm to define areas (or regions) through the application of a concave hull algorithm and classify a given location according to the characteristics of POIs within the cluster. Griffin et al. (2006) use decision trees to classify GPS-derived POIs (i.e., personal locations to a given individual). The main goal of their approach is to automatically classify trips. Clusters of trip-ends are determined using a density-based clustering algorithm, and the generated clusters are classified as "home", "work", "restaurant", etc., based on the time of day and length of the stay. To the best of our knowledge, no previous efforts have been made to classify POIs in a standard system such as the NAICS, although NACIS is widely used for industry classification and has already been used, for instance, to classify websites through machine learning techniques (Pierre, 2001).

2.3. Increasing demand for disaggregated land use data

In the past two decades, accompanied by improved computing power, availability of disaggregated GIS data, and growing interest in learning human economic activities at an increasingly fine-grained spatial level, the demand for collecting disaggregated land use data has increased greatly. Among the great variety of urban and regional research and applications, such examples include large-scale urban simulation models (Batty, 2003; Waddell, 2002; Waddell, Wang, & Charlton, 2008) and activity-based LUTE models (Bowman & Ben-Akiva 2001; Salvini & Miller, 2005; Bradley, Bowman, & Griesenbeck, 2007; Ferreira, Diao, Zhu, Li, & Jiang, 2010). These models have evolved from requesting data at the traditional aggregate level (e.g., census tract or TAZ level) to more disaggregated level (e.g., census block or parcel level). Beyond the fields of transportation and environmental studies, disaggregated land use data (i.e., employment size by category) can also be used to analyze urban economies. For example, Currid and Connolly (2008) try to understand the importance of agglomeration economies as a backbone to urban and regional growth by identifying clusters of several "advanced" service sectors (professional, management, media, finance, art and culture, engineering and high technology) and comparing them among the top ten populous metropolitan areas in the U.S.

Economic activities tend to be more concentrated or clustered than residential locations; therefore, the traditional disaggregation approach—assuming uniform distribution of economic activities across space—is not plausible. However, employment data with detailed size, type and location are expensive and not well understood. For example, with the demise of the U.S. Census long form (which contained additional questions and provided more detailed socioeconomic information about the population[8]), obtaining employment data directly from government-published data sources at the disaggregated level is difficult for years after 2000.

In this study, we attempt to utilize emerging publically available VGI data to develop new data-fusion methods for estimating disaggregated land use, which are more easily restructured as models and conditions change. Detailed POI information can provide high resolution information to support activity-based LUTE models and agent-based urban simulation (especially for work location choice and destination choice models) and can be useful for analyses of urban and regional economies, as our study provides a new method to disaggregate employment sizes into spatial units of a resolution higher than that in which the public data are readily available. Examples of such public data include County Business Patterns (CBP) or ZIP Code Business Patterns (ZBP) data. These datasets provide the number of establishments by employment-size classes and by detailed industries in the U.S.[9] but at much coarser geospatial levels.

3. Overview of research framework and study area

As discussed above, the main purpose of this study is to develop and test a new method for estimating land use (i.e., employment size by category) at a disaggregated level for metropolitan areas by using VGI—POI data. The overall structure for this study is illustrated in Fig. 1. We proceed in two stages. (1) By using the machine learning method and a training set of proprietary POI data (D&B data), we automatically classify online POI (Yahoo! data) into standardized NACIS categories. By combining the classified POI information (on industrial classification, location, and others) with aggregate employment data (at the US census block group level), we then estimate employment sizes by category at the disaggregated level (the US census block level). (2) Following the same disaggregation methodology, we combine POIs obtained from another independent proprietary business establishment database (i.e., infoUSA) with aggregate census employment data (at the census block group level) to disaggregate land use as a benchmark for ground truth.

We focus on an area in which all data (e.g., online POIs, GIS, census employment data, and business establishment data) required to develop, calibrate, and validate the proposed new model are available. We select 6 towns located within the first ring road (Route 128) in the Boston metropolitan area (Fig. 2) as an example. 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 area, population and employment sizes and densities in these 6 towns.

Due to the substantial efforts of MassGIS (the Commonwealth's Office of Geographic and Environmental Information), ample GIS data for the Boston metro area are available for public use. In addition, we utilize two sets of proprietary business establishment data (i.e., D&B and infoUSA data) for the Boston metro area for model training and validation, respectively. This new approach will help derive disaggregated land use estimations (measured by employment size by category) and facilitate urban modeling efforts undertaken by local agencies.

Detailed data sources for this study include the following:

- 64,133 POIs from Yahoo! for the Boston metropolitan area within the first ring road (Route 128) of the metro area for all categories;
- 29,402 POIs for all categories from the 2007 D&B database for the same area;
- Employment sizes by category at the census block group level obtained from the 2000 Census Transportation Planning Products (CTPP) database;
- GIS data for the boundaries of towns, block groups, and blocks obtained from MassGIS public online data sources;
- 2008 infoUSA business establishment data, which are used for model evaluation in Section 5.

4. Machine learning: automatically classifying online POIs

Due to their nature, online POI databases usually grow faster than proprietary POI databases such as D&B or infoUSA business establishment databases. However, there often exist duplicated POIs in online user-content sources, and their categorization does not follow a standardized classification system (such as the NAICS) that is used in most proprietary business establishment databases. We hypothesize that there is considerable coherence between categories of online VGI platforms (e.g., Yahoo!) and NAICS codes; therefore, a model could be trained to automatically classify incoming online-extracted (e.g., Yahoo!) POIs.

4.1. Official taxonomies

In business, classification systems serve to communicate important facts about companies as shorthand for users interested in a particular area of industry or a specific business sector (Hodge, 2000). The NAICS, the International Standard Industrial Classification (ISIC) (United Nations, 2012), and the Classificação de Actividades Económicas (Economic Activities Classification – CAE) (CAE, 2012) are examples of official and standard POI classification systems. All responsible entities for these classification systems provide a complete listing and mapping of categories online. Coding systems usually group industries in a hierarchy, with the

[8] Source: https://www.census.gov/history/www/programs/demographic/american_community_survey.html (Last retrieved in August, 2014.)

[9] Source: https://www.census.gov/econ/cbp/ (Last retrieved in August, 2014).

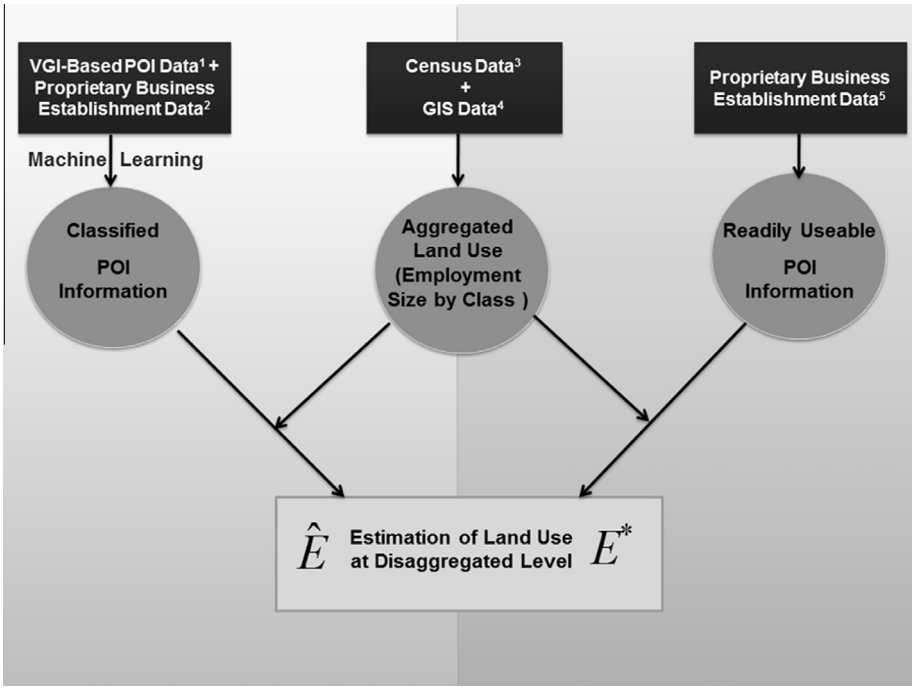

Fig. 1. Overview of research framework. *Notes:* 1. The VGI-based POI data in this study are obtained from Yahoo! 2. The proprietary business establishment data set for training and classifying the VGI-based POIs in this study is the D&B data set. (The choice of D&B or infoUSA should have no impact on the POI classification results, assuming these two sources give equally good information on POIs.) 3. CTPP 2000 data on employment at the census block group level was used in this study. 4. GIS data used here include the boundaries of spatial analysis units (such as towns, census block groups, and census blocks) and are obtained from MassGIS. 5. The proprietary business establishment data set for the purpose of validation (ground truth) in this study is the infoUSA data set. (Again, the choice of infoUSA or D&B should have no impact on the estimation results, as long as the independent proprietary POI data set is different from the training data set for ground truth. We assume these two sources give equally valid POI information.)

"major industry sectors" at the top of the hierarchy and then gradually becoming more specific further down this hierarchy. Although some coding systems have different levels of detail in taxonomy, all systems classify a business establishment by its most profitable activity when in different industrial sectors. In our case, NAICS is the most convenient choice, given the available databases. Fig. 3 shows part of the NAICS hierarchy.[10]

4.2. POI Matching

We use a *POI Matching* algorithm that maps POIs from Yahoo! to D&B to generate training data for the machine learning algorithms.[11] As a consequence, these POIs have both Yahoo! categories and NAICS codes provided by the D&B database. Fig. 4 presents the POI Matching algorithm (Cohen, Ravikumar, & Fienberg, 2003) that we employed to identify similar names, ignoring misspelling errors and some abbreviations. This algorithm[12] consists of a set of rules to compare POIs according to their names, websites and geospatial distances. The similarity of POI names is measured between two strings and normalized using 0 to represent no similarity and 1 as an exact match. We set the similarity thresholds to high values to obtain matches with high confidence. By manually validating a random subset of the identified POI matches (6 sets of 50 random POIs assigned to 6 volunteers), we concluded that the percentage of correct similarities identified was above 98% ($\sigma = 1.79$, SE = 0.73). Unlike validations discussed below, this validation is highly objective and does not demand external participants or a very large sample.[13]

We estimate that the category taxonomy for a Yahoo! POI has more than 1300 distinct categories distributed in a 3-level hierarchy; on average, each POI in the Yahoo! database is assigned to roughly two categories. On the other side, the D&B data that we used in this study cover 514 distinct six-digit NAICS codes. While the 2007 NAICS taxonomy has a total of 1175 six-level categories, our sample data in the selected study area only cover the most common NAICS codes.

An analysis of the coherence between NAICS and Yahoo! business categories shows that only 80.2% of POIs have a consistent corresponding NAICS code with the most common one for the same set of categories. For example, both POIs "*Brueggers Bagel Bakery*" and "*Rebeccas Cafe*" belong to the Yahoo! categories "*Cafes; Bakeries; American Restaurants; Sandwiches; Coffee Houses*"; however, the former is classified with the six-digit NAICS code 311811 (*retail bakeries*), while the latter has the NAICS code 813910 (*eating house associations*). Thus, approximately one fifth of the POIs are incoherent with the rest of the sample. This result highlights the problem of allowing users to add arbitrary categories to their POIs without restrictions. For 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 POI cases correctly.

[10] Full NAICS codes have exactly 6 digits, structured in the following manner: the first two digits designate the economic sector, the third digit designates the subsector, the fourth digit designates the industry group, the fifth digit designates the NAICS industry, and the sixth digit designates the national industry (NAICS, 2012).

[11] As mentioned above, the choice of D&B or infoUSA data sets for training should have no impact on the POI classification results, assuming these two sources give equally valid POI information.

[12] The JaroWinklerTFIDF algorithm proposed by Cohen et al. (2003).

[13] Using the central limit theorem, the standard error of the mean should be near 0.73. Assuming an underestimation bias for $n = 6$ of 5%, the accuracy remains very high, yielding a 95% confidence interval of [96.5%, 98.7%].

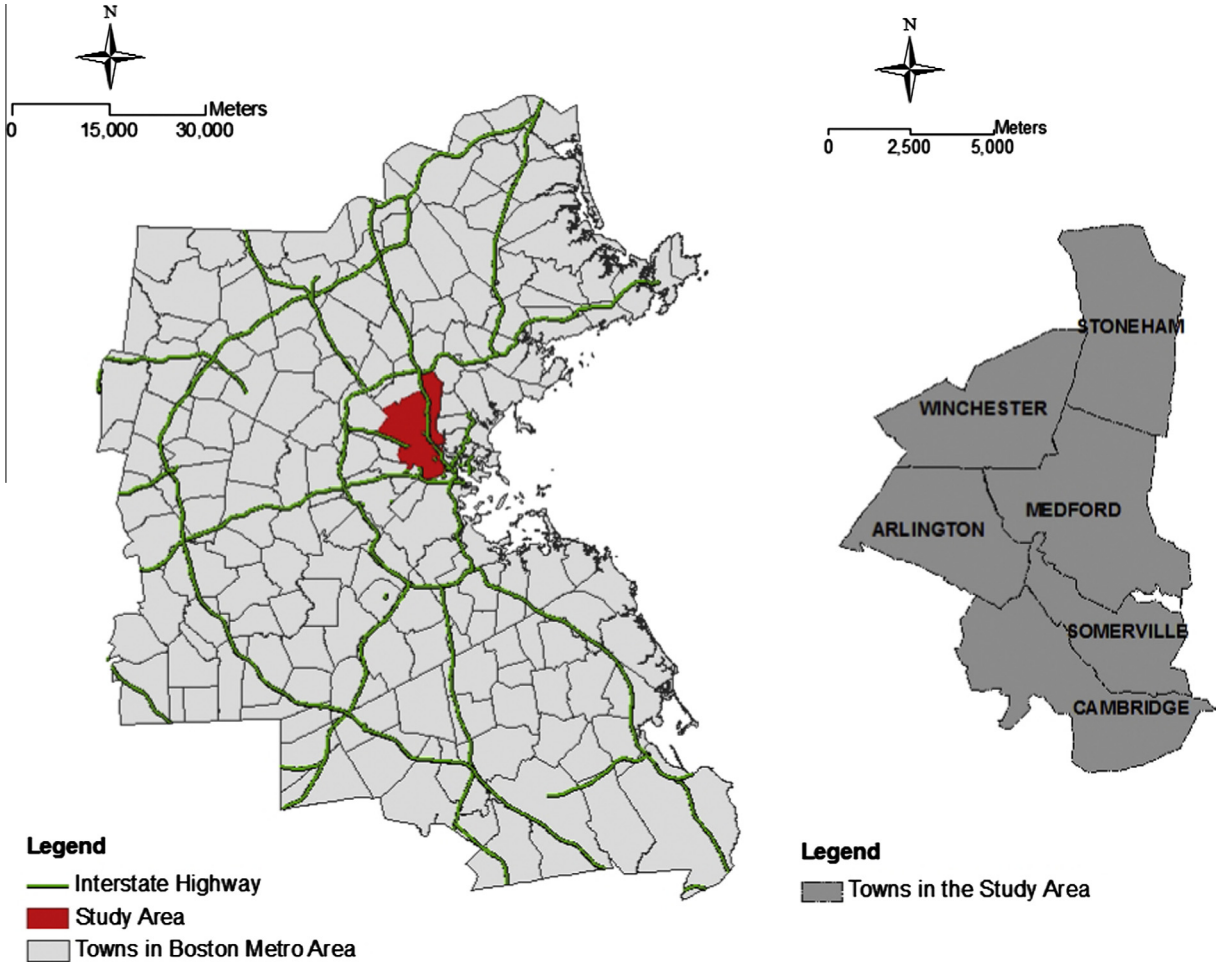

Fig. 2. Boston metropolitan areas and 6 selected towns in the study area.

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

Town name	Population, 2000	Employment, 2000	Area (sq km)	Pop. density (residents/sq km)	Emp. density (workers/sq km)
Arlington	42,389	8577	13.44	3154	638
Cambridge	101,355	114,763	16.66	6084	6889
Medford	55,765	22,071	20.96	2661	1053
Somerville	77,478	22,832	10.65	7275	2144
Stoneham	22,219	8660	15.94	1394	543
Winchester	20,810	7400	15.65	1330	473

Data Source: U.S. Census 2000 and MassGIS.

51 - Information
511 - Publishing Industries (except Internet)
5111 - Newspaper, Periodical, Book, and Directory Publisher
511110 - Newspaper publishers and printing combined
511120 - Periodical Publishers
511130 - Book Publishers
5112 - Software Publishers

Fig. 3. Example of NAICS hierarchy.

4.3. POI classification

After matching POIs from Yahoo! to D&B, we then use Weka (Witten & Frank, 2005), a data mining platform that provides a portfolio of classification algorithms. We show that it is possible to classify POIs to the widely used NAICS system with several different machine learning algorithms using only the categories or tags that are commonly associated with them. We use supervised learning over collected POIs from Yahoo! and compare the results using the proprietary POI (i.e., D&B) data as ground truth. We implement a classification strategy that directly assigns a NAICS code to a POI given its Yahoo! categories. Each NAICS code is simply considered an isolated string "tag" that is assigned to a POI.

In our experiments, we classified POIs for different NAICS levels, particularly two-, four- and six-digit NAICS codes. Two-digit codes allow analysis of economic sectors, while six-digit codes specify the detailed categories of business establishments. We tested different types of machine learning algorithms, namely Bayesian networks, tree-based learners, instance-based learners and rule-based learners. Neural networks or Support Vector Machines are not ideal for this problem due to the high number of classes. While these algorithms can yield good results, their training is impractical.

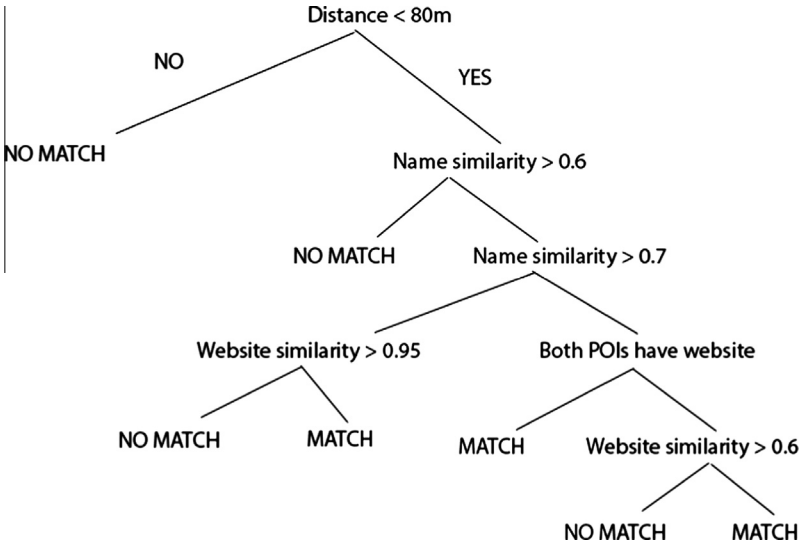

Fig. 4. POI Matching algorithm.

For validation purposes, we use ten-fold cross-validation (Mitchell, 1997), in which we divide the dataset into 10 parts, train with 9 of them and test with the remaining one. We run exactly 10 of these procedures and average the results. 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. Table 2 shows the accuracies obtained using different machine learning algorithms for different NAICS levels (two-, four- and six-digit codes) for extracted POIs. Kappa is a measure of agreement normalized for chance agreement and is computed as $K = (P(A) - P(E))/(1 - P(E))$, where P(A) is the percentage agreement (i.e., between classifier and ground truth) and P(E) is the chance agreement. $K = 1$ indicates perfect agreement, and $K = 0$ indicates chance agreement.

The tree-based (e.g., ID3 and RandomForest) and instance-based learning approaches (e.g., IBk and K^*) perform best in this classification task, especially the latter. Note that at the sixth level, only 80.2% of the NAICS codes in the data were assigned in a totally non-ambiguous way. The most successful algorithm is IBk (with $k = 1$), which essentially finds a similar test case and assigns the same NAICS code. The difference in accuracy between tree-based and instance-based approaches is insufficiently large to conclude which one outperforms the other. However, we could expect that instance-based models yield better results because the distribution of different Yahoo! categories is relatively even among examples of the same NAICS code (implying no clear dominance of some categories over others). Unsurprisingly, the Naive Bayes algorithm performs badly because the assumption that different Yahoo! categories for the same NAICS classification are independently distributed is obviously false—for example, "doctors & clinics", "laboratories", and "medical laboratories" are highly correlated. This assumption is not fully necessary in Bayesian Networks, which actually yield better results. Unfortunately, we could not find a model search algorithm that performs in an acceptable time (less than 72 h) and produces a more accurate model. We ultimately used Simulated Annealing and Hill Climbing.

As expected, we obtained better results classifying POIs with two-level NAICS codes than with six-level NAICS codes because the noise due to ambiguous NAICS codes assignments in the POI dataset is smaller—we now have 87.1% of non-ambiguous cases. Finally, we ran the classifier for the entire Yahoo! dataset to obtain the necessary NAICS codes for the remaining POIs that did not find a match in the D&B dataset.

Table 2
Accuracies obtained by different machine learning algorithms with POIs for the Boston area.

Algorithm	NAICS2 (kappa)	NAICS4 (kappa)	NAICS6 (kappa)
ID3	85.495 (0.842)	77.955 (0.776)	74.015 (0.737)
C4.5	84.241 (0.828)	77.630 (0.772)	73.071 (0.727)
Random Forest	86.174 (0.849)	79.298 (0.789)	74.753 (0.744)
JRip	81.334 (0.795)	74.340 (0.737)	69.264 (0.686)
IB1	82.736 (0.812)	74.266 (0.738)	68.644 (0.683)
IBk (with $k = 1$)	86.646 (0.854)	79.475 (0.791)	75.343 (0.750)
K^*	85.702 (0.844)	79.726 (0.794)	75.387 (0.751)
BayesNet	80.950 (0.790)	56.721 (0.554)	45.064 (0.438)
NaiveBayes	74.399 (0.715)	40.446 (0.382)	30.264 (0.283)

5. Disaggregating land use using POIs

Fig. 5 demonstrates the modeling processes to estimate the employment size and density by category at the disaggregated level (e.g., census block level). The process follows the general framework described in Section 3. We make two assumptions to disaggregate land use (i.e., employment size by category) at the census block level. First, we assume that, for a census block group (which consists of a group of census blocks), the employment size for the same industrial category at each POI is independent and identically distributed (i.i.d.). Second, due to measurement errors in identifying POI locations in the geocoding process (i.e., the error in allocating a census block of which a POI is part), we make a 25-m buffer area around each POI and calculate the probability that a POI is in each census block based on the share of each census block in the buffered area of the POI (this treatment is discussed in detail in Section 5.1.1). Based on these two assumptions, we estimate the employment size by category in each census block by summing the estimated employment size at each POI (which equals the product of the average employment size by category at each POI and the probability of the POI appearing in this census block).

5.1. Data

5.1.1. POIs

For activity-based LUTE analysis, employment size (or density) by category is the most important data input to estimate work location choice and destination choices. These employment categories usually match with the two-digit NAICS codes (which include approximately 20 categories of sectors, such as retail, manufacturing, professional services, educational services, and public

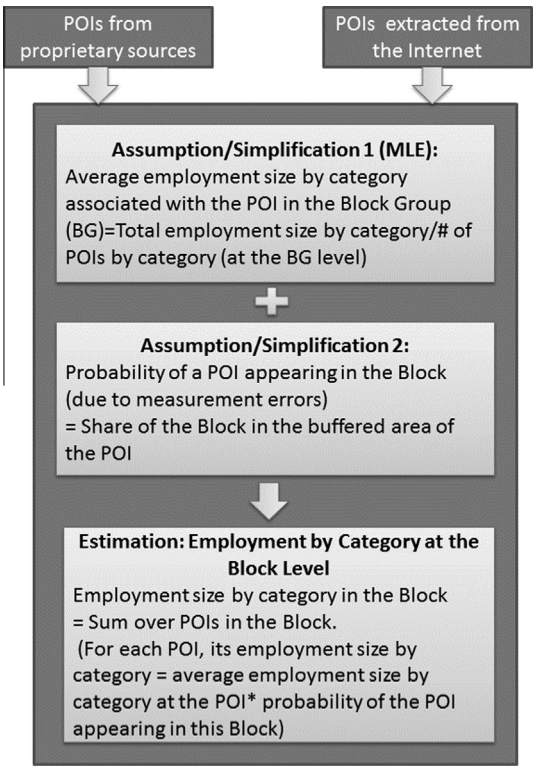

Fig. 5. Model estimation process.

administration). In this section, we focus on POIs in the retail sector (two-digit NAICS code = 44 or 45) as a demonstration due to space limitations. POIs in the retail sector are only a part of all the POIs that we trained and classified in Section 4 for the study area. Fig. 6 displays the retail POIs obtained from Yahoo![14] (left) and infoUSA (right). Tables 3 and 4 summarize the number of POIs by 3-digit NAICS category and by town. The online-extracted (Yahoo!) POIs identify approximately 98% POIs of the proprietary (infoUSA) source, and this ratio varies across categories and towns.

One threat to data validity comes from the geographic information of the POIs. First, our proposed method of estimating disaggregated employment size depends heavily on the geocoded locations of POIs. However, in most cases, points with *X*/*Y* coordinates are usually geocoded 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 geocoding errors. Thus, systematic measurement errors may exist within the same source and/or across different sources. Incorporating methods that can reduce this type of errors is very important to the reliability of this study. To address the problem of potential geocoding errors, we create a buffer area with a 25-m radius from each POI and use the area share of each block in the POI buffer area as the probability that each POI may exist in that block. The 25-m size is determined by the relative road width and block size—the buffer size must be sufficiently large to cover both sides of the road but not too large to cover the entire block at each side.

5.1.2. Aggregate retail employment data

The choice of spatial analysis unit at the aggregate level (e.g., transportation analysis zone, census tract, or census block group) depends on the availability of data and estimation accuracy concerns. For example, employment-by-category data for our study 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. As the census block group (BG) has higher spatial resolution than the TAZ level, we disaggregate the census employment data from census block group level to census block level by using the extracted online POIs (from Yahoo!) as discussed in Fig. 5. The CTPP database distinguishes 14 major categories of employment (e.g., agriculture, construction, manufacturing, wholesale, retail, transportation, information industry, finance industry, professional services, educational industry, recreation and food service industry). These 14 employment categories in the CTPP database have a one-to-many relationship with the NACIS two-digit codes. For example, the retail category in the CTPP database corresponds to NAICS two-digit codes of 44 or 45.

Fig. 7 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 equal because we have not yet used POI information to differentiate the blocks within a block group. Table 5 describes the numbers of block groups and blocks in the 6 selected towns.

5.2. Estimating disaggregated land Use

5.2.1. Estimation method: maximum likelihood estimation (MLE)

We use a set of POIs extracted and classified from a user-content platform (i.e., Yahoo!) to disaggregate the aggregate data to a finer level and use infoUSA, which contains detailed information of business establishments in the United States, to evaluate our newly developed method. To support LUTE modeling, in which travel demand is sensitive to block level travel time and distances, we would like to have land use identified 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 land use 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 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. Because the *X*/*Y* location of POIs includes measurement error, we buffer the *X*/*Y* locations and treat the assignment of POIs to blocks as a random variable.

5.2.2. Estimation evaluation

By employing the MLE method described above and proprietary business establishment data (e.g., infoUSA data[15]), we obtain a benchmark employment size of category c at block b in block group g, $E^{*}_{b,c,g}$, which is considered the true value of the disaggregated employment size. By using the derived VGI POI information tagged with NAICS, we obtain an ML estimate of employment size of category c at block b in block group g, $\widehat{E}_{b,c,g}$. 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. To compare our method with the traditional disaggregation approach (assuming spatially uniform distribution of employment opportunities), we use the ratio of MSEs of our MLE method and the traditional uniform disaggregation method,

[14] It is worth reminding the reader that the NACIS code for these Yahoo! POIs are trained by using the other proprietary POI data source (D&B), as discussed in Section 4.

[15] Again, the choice of infoUSA or D&B should have no impact on land use disaggregation estimation, by assuming these two sources give equally valid POI information. However, we have used D&B POI data to classify Yahoo! POI categories, so we avoid using D&B POI data again for ground truth purpose.

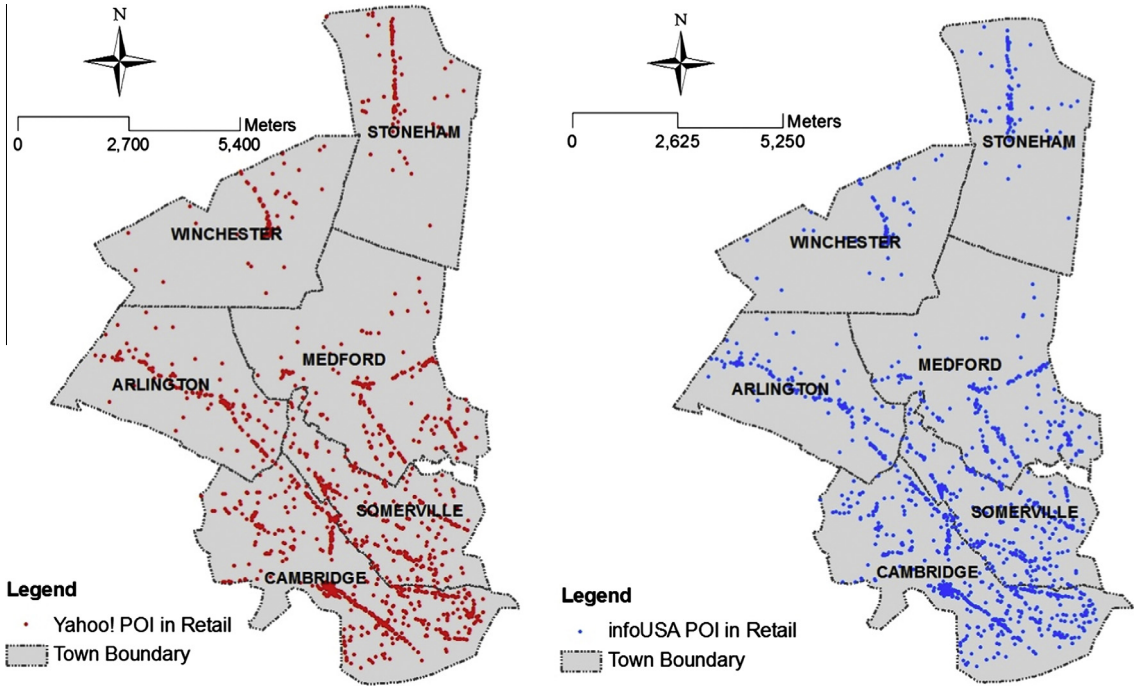

Fig. 6. Distribution of retail POIs from Yahoo! (left) and infoUSA (right) in the study area.

Table 3
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 (%)
441	Motor vehicle and parts dealers	96	99	103.13
442	Furniture and home furnishings stores	104	113	108.65
443	Electronics and appliance stores	251	265	105.58
444	Building material and garden equipment and supplies dealers	104	123	118.27
445	Food and beverage stores	268	209	77.99
446	Health and personal care stores	130	164	126.15
447	Gasoline stations	79	90	113.92
448	Clothing and clothing accessories stores	267	228	85.39
451	Sporting goods, hobby, book, and music stores	175	191	109.14
452	General merchandise stores	61	30	49.18
453	Miscellaneous store retailers	301	285	94.68
454	Non-store retailers	17	26	152.94
Total		1853	1823	98.38

Table 4
Statistics of retail POIs in the study area from Yahoo! and infoUSA by town.

Town name	infoUSA Count	Yahoo! Count	Yahoo! to infoUSA (%)
Arlington	174	188	108.05
Cambridge	830	816	98.31
Medford	301	292	97.01
Somerville	340	338	99.41
Stoneham	113	126	111.50
Winchester	93	86	92.47

the *relative mean squared error* (RMSE), to evaluate the goodness of fit of our model.

Fig. 8 shows the estimation results of the disaggregated retail employment density at the census block level in the 6 towns of our study area, using proprietary POI data (infoUSA) and VGI-based POI data (Yahoo!). By comparing the estimation results, we find that the disaggregated employment estimations using these two different POI data sources are very similar.

We use the RMSE to quantitatively evaluate the goodness of fit of the model; the rigorous mathematical form of RMSE is given in Eqs. (1) and (2).

$$\overline{E}_{b,c,g} = \frac{w_{b,g} E^{*}_{c,g}}{\sum_{q} w_{q,g}} \tag{1}$$

$$\mathrm{RMSE}\left(\widehat{E}, E^{*}\right) = \frac{\sum_{b,c,g}\left(\widehat{E}_{b,c,g} - E^{*}_{b,c,g}\right)^{2}}{\sum_{b,c,g}\left(\overline{E}_{b,c,g} - E^{*}_{b,c,g}\right)^{2}} \tag{2}$$

In Eq. (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 that assumes that employment is spatially uniformly distributed across blocks in each block group g. In Eq. (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 (e.g., infoUSA[16]); $\widehat{E}_{b,c,g}$ is the maximum likelihood estimate of employment size of category c at block b in block group g, employing the online-extracted Yahoo! POIs.

The RMSE is the ratio of the MSE of the disaggregated land use

[16] Please refer to footnote 15 for discussion of the use of infoUSA POI data instead of D&B POI data as the benchmark.

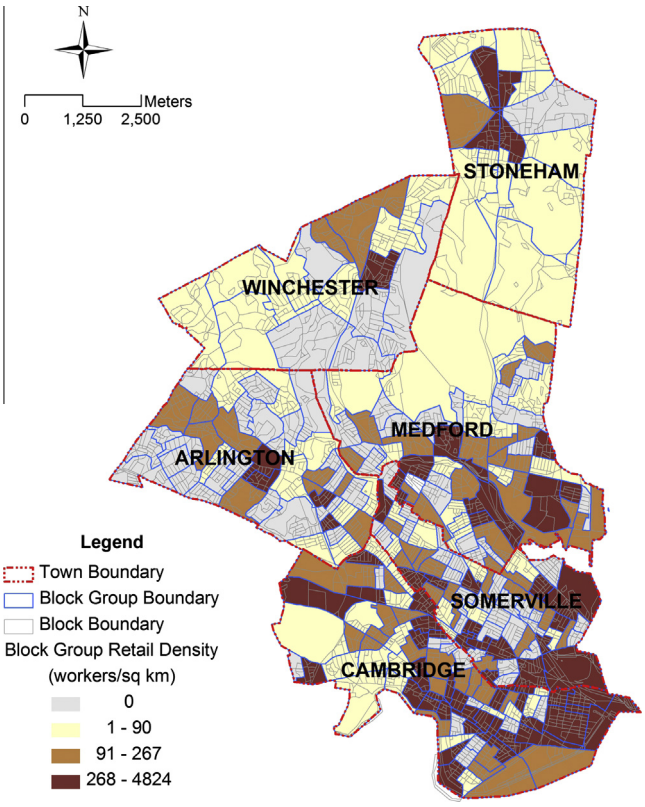

Fig. 7. Aggregated employment densities at the block group level.

estimate using POI information to the MSE using the traditional block group average estimation method. If the RMSE is less than 1, then our new method using the derived POIs improves land use disaggregation; lower RMSE values indicate greater improvements by the data-fusion method. If the RMSE is close to 0, then the method using online-extracted POIs gives very similar estimates as those obtained from the proprietary POI database. However, if the RMSE is greater than 1, then the derived POIs do not well reflect the distribution of population POIs (as listed in the proprietary business establishment database). As described in the POI data description section, our online-extracted POIs (from Yahoo!) do not match perfectly with the proprietary business establishment data (from infoUSA). However, we conjecture that, on average, the estimations of disaggregated employment at the block level will be improved compared to the traditional uniform disaggregation approach; to some degree, these POIs represent the distribution of economic activities across space and reflect their heterogeneous nature. Employing Eq. (2), the disaggregated employment estimation at the block level using Yahoo! POIs gives RMSE = 0.309. The RMSE is significantly smaller than 1, which means that using the online-extracted Yahoo! POIs to estimate disaggregated employment sizes at the block level has reduced the mean squared error by approximately 70% 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. The impacts of POI geocoded errors in blocks with large areas are relatively smaller than those in blocks with smaller areas; the relative gaps between the online-extracted 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 geocoded errors matter less (because street width is a small fraction of block size). We sorted the 3633 blocks (with complete data within our study area) by their areas and divided them into two groups—one consisting of 1817 blocks with smaller areas and the other consisting of 1816 blocks with larger areas. We computed the RMSE for each group; the RMSE for the group with smaller block sizes is 0.432, and the RMSE for the group with larger block sizes is 0.299. These results are consistent with our conjecture.

Table 5
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: U.S. Census 2000 and MassGIS.

6. Conclusions

According to our case study, by using volunteered geographic information in the form of points of interest together with publicly available aggregate employment data from the census at the aggregate level, we can derive more accurate land use estimations at the disaggregated level than the traditional disaggregation approach, which assumes uniform distribution of land use across the aggregate spatial level. In general, extracted online POI data are very useful in estimating accurate disaggregated land use, although there are several issues of data validity and reliability. First, because POI information is extracted from online platforms, the coverage and accuracy of the 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 has been widely used to provide and acquire information, the POI information can be widely accessible, but potential gaps may exist between the total business establishments and the available information online. These gaps can be reduced as more cities improve their information technology infrastructure and online user-content platforms for publishing POI information apply rigorous and standardized categorization guidelines. Combining different online sources can help reduce these gaps but may also introduce problems of redundancy, which has been addressed by our machine learning method. However, the issue of data currency may not be easily adjustable because VGI-based POI data may not reflect obsolete business establishments instantaneously, even though these POIs may disappear with time.

Depending on the requirements of different types of urban analyses, the frequency of obtaining updated VGI-based POI data can vary as well. For example, researchers can use annually updated VGI-based POI data to estimate disaggregated land use for LUTE model each year. The cost of obtaining these online POIs can be very low once the machine learning classification method has been developed, as the same algorithm can be run repeatedly for updated POI data. However, with various degrees of restrictions on the massive usage of the VGI data according to legal terms by different VGI service providers, there may be additional costs associated with obtaining VGI data.

In general, for cities without resources to purchase or update proprietary business establishment data (e.g., infoUSA data), our methods presented here provide an alternative to developing timely disaggregated land use estimations, which are essential for activity-based LUTE models. In a separate study, we have

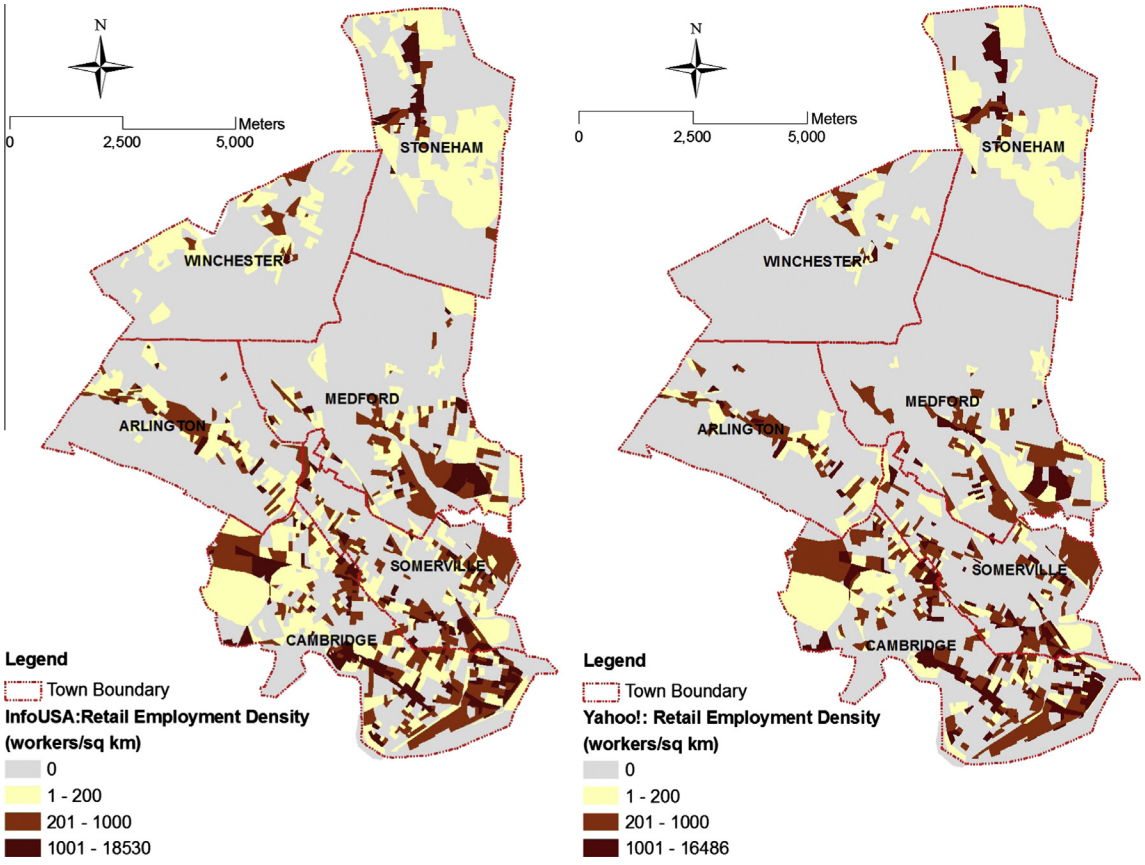

Fig. 8. Disaggregated retail employment density at block level using infoUSA (left) POIs and Yahoo! (right) POIs.

applied this disaggregation approach (using extracted online POIs and aggregate employment data) to Lisbon, Portugal, in an integrated LUTE model developed by the MIT-Portugal program (MPP). Disaggregated land use information is very important to improve travel demand models, partially because travel demand is very sensitive to micro-level changes in travel time and distances. As destinations tend to be more clustered and concentrated than residential locations, the location and categorization information of POI is very useful for planners to understand characteristics and derived travel demand at the micro-level. Meanwhile, as discussed in previous sections, the classification and disaggregation of land use using VGI-based POI can also be very useful for analyzing urban and regional economies, especially when disaggregated employment data by size and industrial category will not be easily available in the future at the local level from public sources, such as the US County Business Patterns or ZIP Code Business Patterns data. The methods developed in this study will be helpful in facilitating researchers and planners to study micro-level travel behavior, travel demand, and urban economies and will create new opportunities for cities with limited resources that wish to develop policy-sensitive urban models at the disaggregated level.

Acknowledgements

We acknowledge partial support from the Singapore National Research Foundation (NRF) through the "Future Urban Mobility" program of the Singapore-MIT Alliance for Research and Technology, and from the Fundação para a Ciência e a Tecnologia (FCT) through the MIT-Portugal Program and the Grant PTDC/ECM-TRA/1898/2012 (INFOCROWDS).

References

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.

Berjani, B. & Strufe, T. (2011). A recommendation system for spots in location-based online social networks. In *Proceedings of the 4th workshop on social network systems*, SNS '11 (pp. 4: 1–4: 6). New York, NY, USA. ACM. ISBN 978-1-4503-0728-4.

Bowman, J., & Ben-Akiva, M. (2001). Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A, 35*, 1–28.

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.*

CAE (2012). *Código de Actividades Económicas*. Instituto Nacional de Estatística. <http://webinq.ine.pt/public/files/inqueritos.aspx?Id=101> Last visited: December, 2012.

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. (2011). Exploring millions of footprints in location sharing services. In *ICWSM '11*.

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 the Web* (IIWeb-03), Acapulco, Mexico.

Cranshaw, J., Hong, J. I., & Sadeh, N. (2012). The Livehoods Project: Utilizing social media to understand the dynamics of a city. In *The Sixth international AAAI conference on weblogs and social media* (pp. 58–65).

Currid, E., & Connolly, J. (2008). Patterns of knowledge: The geography of advanced services and the case of art and culture. *Annals of the Association of American Geographers*, 414–434.

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers, 102*(3), 571–590. http://dx.doi.org/10.1080/00045608.2011.595657.

Ferreira, J., Diao, M., Zhu, Y., Li, W., & Jiang, S. (2010). Information infrastructure for research collaboration in land use, transportation, and environmental planning. *Transportation Research Record: Journal of the Transportation Research Board, 2183*, 85–93.

Griffin, T., Huang, T., & Halverson, R. (2006). Computerized trip classification of GPS data. In *Proceedings of 3rd international conference on cybernetics and information technologies, systems and applications* (CITSA 2006) (pp. 22–30).

Hodge, G. (2000). Systems for knowledge organization for digital libraries: Beyond traditional authority files. *Technical report, Digital Library Federation*, April 2000. <http://www.clir.org/pubs/reports/pub91/contents.html> Last visited: December, 2012.

Long, X., Jin, L., & Joshi, J. (2012). Exploring trajectory-driven local geographic topics in foursquare. In *Proceedings of the 2012 ACM conference on ubiquitous computing – UbiComp* '12 (pp. 927). http://dx.doi.org/10.1145/2370216.2370423.

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

NAICS (2012). North American Industry Classification System. <http://www.naics.com> Last visited: December, 2012.

Noulas, J., Scellato, S., Mascolo, C., & Pontil, M. (2011). An empirical study of geographic user activity patterns in foursquare. In *ICWSM '11*.

Pierre, J. (2001). On the automated classification of web sites. *Linkoping Electronic Articles in Computer and Information Science, 6*, 1–12.

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.

Santos, M. & Moreira, A. (2006). Automatic classification of location contexts with decision trees. In *CSMU-2006: Proceedings of the conference on mobile and ubiquitous systems, Guimares, Portugal* (pp. 79–88).

Toole, Jameson L., Ulm, Michael, González, Marta C., & Bauer, Dietmar (2012). Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD international workshop on urban computing (UrbComp '12)* (pp. 1–8). New York, NY, USA: ACM.

United Nations (2012). International Standard Industrial Classification of all economic activities, 2011. <http://unstats.un.org/unsd/cr/registry/isic-4.asp> Last visited: December, 2012.

Waddell, P. (2002). UrbanSim: Modeling urban development for land use, transportation and environmental planning. *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. In *TRB 87th annual meeting compendium of papers DVD*. Washington, D.C.: TRB.

Wang, L., Waddell, P., & Outwater, M. L. (2011). Incremental integration of land use and activity-based travel modeling. *Transportation Research Record: Journal of the Transportation Research Board, 2255*(1), 1–10.

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

Yuan, Jing, Zheng, Yu, & Xie, Xing (2012). Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)* (pp. 186–194). New York, NY, USA: ACM.