ARTICLE IN PRESS

Computers, Environment and Urban Systems xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Semantic enrichment of building data with volunteered geographic information to improve mappings of dwelling units and population

Carola Kunze, Robert Hecht *

Leibniz Institute of Ecological Urban and Regional Development Dresden (IOER), Weberplatz 1, 01217 Dresden, Germany

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Building model
Disaggregation
Census data
Cadastre data
Land use
Geographic information system
Population mapping
Dasymetric mapping
Volunteered geographic information

ABSTRACT

Small-scale data on dwellings and population density are required for precise geospatial urban modelling. Further, knowledge of building usage is necessary to model socio-economic aspects such as the distribution of dwellings and population. In an effort to limit costs and resourcing efforts, users and institutes in research and spatial planning are developing strategies to extract such information from existing geographic base data. Currently, land-use information from official datasets merely distinguishes residential from non-residential building usage, but cannot identify areas of non-residential usage inside residential buildings. Additional data sources are therefore needed to fill this gap. In this paper we propose an approach to process semantic information from user-generated OpenStreetMap (OSM) data to specify non-residential usage in residential buildings. This estimation is based on OSM attributes, so-called *tags*, which are used to define the extent of non-residential usage. Our objective is to identify the potentials and reveal the limitations of integrating semantic OSM data for the evaluation of building usage. Official statistical data on dwellings and population is used to validate results. Thereby we prove the benefit of integrating OpenStreetMap semantic data to refine the estimation of non-residential floor area in the study area of the German City of Dresden, Saxony.

1. Introduction

The fields of urban and regional planning as well as ecological surveys and disaster management require detailed information about the number and location of dwellings and their inhabitants (see Ahola, Virrantaus, Krisp, & Hunter, 2007; Aubrecht, Steinnocher, Hollaus, & Wagner, 2009; BBK, 2010). Usually this data can be taken from census datasets or regional surveys. For example, new population statistics for Germany were recently released as part of the register-based census for the year 2011[1]. Demographic and socioeconomic data for European countries and regions are provided by Eurostat. However, census data has a number of drawbacks, such as lengthy periods between data acquisition, the aggregation on large spatial units and the fact that administrative borders can change over time. In some countries census population data can also be of poor spatial resolution (ward level, state level) or non-existent (Alahmadi, Atkinson, & Martin, 2013).

For these reasons, methods have been developed to disaggregate socio-economic information such as population and dwellings from various data sources, whether digital topographic databases or remote sensing imagery (Lo, 1995; Wu, Qiu, & Wang, 2005; Wurm & Taubenböck, 2010). In this way it is possible to determine population distribution with the help of detailed information at the level of buildings. Although remotely sensed data can be used for automatic building extraction and the identification of land use classes (residential or non-residential building), this approach is prone to errors (Wurm & Taubenböck, 2010). Current topographic databases usually contain single building footprints and usage information, in which distinctions are made between residential, industrial/commercial and other uses for leisure, sports, etc.

However, data on building usage generally refers to an entire building or a building block, and does not provide any information on non-residential usage *within* residential buildings (ground-floor shops, offices, etc.). According to Meinel, Hecht, and Herold (2009), this lack of data on non-residential usage affects population mapping and the analysis of settlements structure, leading to over- or underestimations of the number of dwelling units or residents, especially in the city centre.

Other data sources may remedy the deficit of small-scale information on building use. Commercial registers or workplace statistics are possible input data to estimate the floor area used for non-residential purposes, often called non-residential floor space. However, most of these sources are not readily available or are

* Corresponding author. Tel./fax: +49 (0)3514679248.
E-mail addresses: carola.kunze@googlemail.com (C. Kunze), r.hecht@ioer.de (R. Hecht).

[1] https://www.zensus2011.de.

http://dx.doi.org/10.1016/j.compenvurbsys.2015.04.002
0198-9715/© 2015 Elsevier Ltd. All rights reserved.

prohibitively expensive for a majority of potential users. One promising alternative approach is to exploit Volunteered Geographic Information (VGI) (Goodchild, 2007) such as from the free and internationally available OpenStreetMap (OSM). As well as extensive geometric data, OSM offers a variety of semantic information, also describing the building stock.

In this research paper we intend to discuss the potential of semantic OSM data to assist in the estimation of non-residential floor area by identifying non-residential usage within residential buildings. To this end we have developed an estimation model based on official building polygons and OSM point, line and polygon features. This non-residential floor space estimation model is validated by comparing calculated data on dwellings and population to a reference dataset.

2. State of the art

Models of the urban environment require precise data on land use. There already exist several studies in geospatial science which have, for example, undertaken micro-spatial analyses of the urban structure as well as looking at principles of geometrical and semantic data integration. In order to give an overview of this research and subsequently to validate our results, previous works on dwelling and population estimation are presented in the following. In Section 2.3 we give a brief overview of the principles of VGI data collection.

2.1. Micro-spatial analysis of the urban structure

Detailed information on the functional, morphological, and socio-economic structure of the built environment is required for urban modelling. Clearly the building stock is the most important component as it directly affects urban structures such as urban form, building density, or the distribution of dwellings and population. Although some major cities provide 3D-city models, there is a lack of fine-scale data at the level of individual buildings (e.g. information on height, the number of floors, building type, building functions), particularly for smaller towns and rural municipalities. Homogeneous datasets are also needed to enable comparative regional or national studies. Therefore, during the last few years, various methods have been developed to classify and describe urban structures based on small-scale settlement indicators.

One way to capture urban structure is to make use of remote sensing data such as very high resolution (VHR) satellite imagery (Bauer & Steinnocher, 2006; Dogruso & Aksoy, 2007; Geiß et al., 2011; Herold, Liu, & Clarke, 2003; Walde, Hese, Berger, & Schmullius, 2013) or aerial imagery (Banzhaf & Höfer, 2008). More recently, laser-scanner data (LiDAR) has become increasingly applied to urban modelling as it permits the construction of 3D models of buildings (Wurm, Taubenböck, Roth, & Dech, 2009). Meinel et al. (2009) and Hecht, Herold, Meinel, and Buchroithner (2013) have described a method to derive urban structure types from scanned topographic maps at scale 1:25 k. Today topographic vector data provided by National Mapping and Cadastral Agencies (NMCA) is widely used to determine urban structure. Yet while these geo-topographic databases usually contain digital 2D building footprints, they often lack information on building types. There have been several attempts to classify building footprints automatically according to a building typology (Colaninno, Cladera, & Pfeffer, 2011; Henn, Römer, Gröger, & Plümer, 2012; Lüscher, Weibel, & Burghardt, 2009; Steiniger, Lange, Burghardt, & Weibel, 2008). Address point data has also been used to enhance the quality of building classification (Orford & Radcliffe, 2007; Smith & Crooks, 2010).

2.2. Dwelling and population mapping at the level of individual buildings

GIS applications such as disaster management, risk and vulnerability assessment, facility management, public health and planning require precise spatial analysis of dwelling units and population at the scale of individual buildings. Due to concerns over privacy, census data at the building level is not available for public usage. Depending on the specific national policy, such data is aggregated to administrative units (municipalities, district, federal states) or census tracts. In Germany, national statistics on population are available at municipal level only, whereas other European countries such as Austria, Denmark, Estonia, Finland, Ireland, Norway or Slovenia already provide population data in raster format with cell size as small as 100 or 125 m (Eurostat, 2012). For many applications, aggregated data at administrative level does not adequately represent the underlying population distribution.

To resolve this problem, techniques of dasymetric mapping are used to transform census population data into finer map units by means of ancillary data (Mennis, 2003). The technique refers to the process of disaggregation, discussed in detail by Eicher and Brewer (2001) and Maantay, Maroko, and Herrmann (2007). The simplest approach to disaggregation is the areal weighted transformation of population from a source unit to a target unit assuming that the population is proportionate to area (Maantay et al., 2007). However, since population is typically non-uniformly distributed within a unit due to different urban densities, auxiliary data is introduced to distinguish habited from uninhabited land, excluding the latter from the population mapping process. Data on land use/land cover or remote sensing imagery has been employed as such auxiliary sources of data. For example, Gallego, Batista, Rocha, and Mubareka (2011) distribute the population to a 100 m density grid by means of CORINE Land Cover data.

Many micro-spatial analyses such as for local disaster management and urban planning require reliable population data at the level of urban block, plot or building. In the past, GIS expert systems have been developed to disaggregate census data using cadastral data (Maantay et al., 2007) or digital topographic maps (Meinel et al., 2009). These approaches rely on the fact that population is closely linked to the characteristics of the built environment such as the function, size, and height of buildings. First attempts to estimate the population of individual buildings have been undertaken by Lwin and Murayama (2009), using residential building footprints as the target unit for disaggregation. They introduced an areametric and a volumetric method, employing the building footprint area and the number of floors as input information, respectively. Ural, Hussain, and Shan (2011) have modified these areametric and volumetric models by implementing a weighting scheme that differentiates between "houses" and "apartments". A more detailed analysis of building types has been carried out by Meinel, Hecht, and Herold (2008) and Meinel et al. (2009), in which building footprints are extracted from topographic maps; average dwelling and population densities for building footprints are empirically determined for eight different residential building types using census data at block level.

The characterization of buildings in terms of their function, size, and number of floors plays a key role in population mapping. Previous studies have looked at the impact of non-residential usage within residential buildings in distorting the mapping of dwelling unit and population (Lwin & Murayama, 2009; Meinel et al., 2009). In particular, populations are overestimated for buildings in the central districts of urban agglomerations and along main shopping streets (Wurm & Taubenböck, 2010). In such areas

Please cite this article in press as: Kunze, C., & Hecht, R. Semantic enrichment of building data with volunteered geographic information to improve mappings of dwelling units and population. *Computers, Environment and Urban Systems* (2015), http://dx.doi.org/10.1016/j.compenvurbsys.2015.04.002