ARTICLE IN PRESS

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

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Identifying residential neighbourhood types from settlement points in a machine learning approach

Warren C. Jochem*, Tomas J. Bird, Andrew J. Tatem

WorldPop, Department of Geography and Environment, University of Southampton, UK
Flowminder Foundation, Stockholm, Sweden

ARTICLE INFO

Keywords:
Urban morphology
Land use
Texture
Point pattern analysis
Machine learning
Big data

ABSTRACT

Remote sensing techniques are now commonly applied to map and monitor urban land uses to measure growth and to assist with development and planning. Recent work in this area has highlighted the use of textures and other spatial features that can be measured in very high spatial resolution imagery. Far less attention has been given to using geospatial vector data (i.e. points, lines, polygons) to map land uses. This paper presents an approach to distinguish residential settlement types (regular vs. irregular) using an existing database of settlement points locating structures. Nine data features describing the density, distance, angles, and spacing of the settlement points are calculated at multiple spatial scales. These data are analysed alone and with five common remote sensing measures on elevation, slope, vegetation, and nighttime lights in a supervised machine learning approach to classify land use areas. The method was tested in seven provinces of Afghanistan (Balkh, Helmand, Herat, Kabul, Kandahar, Kunduz, Nangarhar). Overall accuracy ranged from 78% in Kandahar to 90% in Nangarhar. This research demonstrates the potential to accurately map land uses from even the simplest representation of structures.

1. Introduction

As populations around the world become more urbanised, particularly in developing countries, the ability to quantify and study the growth and changing function of cities in detail has become more important for urban growth, informal settlements, poverty, environmental and health concerns (Duque, Patino, Ruiz, & Pardo-Pascual, 2015; Herold, Liu, & Clarke, 2003; Kuffer, Pfeffer, & Sliuzas, 2016; Kuffer, Pfeiffer, Sliuzas, & Baud, 2016; UN Habitat, 2016). Moreover, the Sustainable Development Goals (United Nations, 2014) and the New Urban Agenda (United Nations, 2017) have brought additional focus for policymakers on land use planning to create resilient, sustainable, and inclusive cities. To meet such goals, data on intra-urban differences in land uses is needed. Yet the speed of population growth and urbanisation makes it necessary to explore new approaches to assist in producing timely and accurate data on cities and regions.

Recent work to identify land use types across large urban areas has increasingly made use of advances in very high spatial resolution satellite or aerial imagery (Cheriyadat, 2014; Kuffer & Barros, 2011). Similar analyses using large collections of geospatial vector data (points, lines, polygons) have received far less attention in the literature than remote sensing approaches, though several studies have noted the potential to identify classes of buildings or urban land uses (Barr, Barnsley, & Steel, 2004; Hecht, Meinel, & Buchroithner, 2015; Longley & Mesev, 2000; Steiniger, Lange, Burghardt, & Weibel, 2008). Classifying land uses, whether based on imagery or vector data, all rely on the assumption of linking observed spatial forms with different functions or land uses on the ground (Barr et al., 2004). The method developed here uses an existing vector dataset of points representing dwellings (referred to here as settlement points) and applies various measures to quantify the multi-scale, spatial patterns to establish that link and train a machine learning algorithm. The goal is to identify areas of different settlement types and to predict those types into unmapped areas. Land use polygon features (Government of the Islamic Republic of Afghanistan [GoIRA] & UN Habitat, 2015) provide training data for a two-class typology of regular and irregular housing. We then describe several metrics calculated from the spatial point patterns of settlement points which are used to characterise the density and distribution of settlements. While the point geometry-derived features alone provide remarkable accuracy in predicting these classes in our case study of seven provinces in Afghanistan, incorporating additional measures of vegetation, elevation, slope, and nighttime lights improves overall accuracy of the classification, reaching up to 90% accuracy. We identify several spatial relationship measures and scales that were most effective

---

* Corresponding author at: WorldPop at University of Southampton, Geography and Environment, Cumberland House, Grosvenor Square, Southampton, SO15 2BG, United Kingdom.
E-mail address: w.c.jochem@soton.ac.uk (W.C. Jochem).

https://doi.org/10.1016/j.compenvurbsys.2018.01.004
Received 13 September 2017; Received in revised form 6 December 2017; Accepted 7 January 2018

Please cite this article as: Jochem, W.C., Computers, Environment and Urban Systems (2018), https://doi.org/10.1016/j.compenvurbsys.2018.01.004

in differentiating residential classes in our study area. Our methodology contributes to ongoing developments in computational methods for utilising big geospatial data, and we implement a variable selection algorithm to select from a large number of correlated features. This work also demonstrates several paths forward for future research. Pattern analysis of vector data can be combined with other remotely sensed data to enhance analyses, as we show with several commonly available satellite-derived measures. Overall the results suggest that our method has potential to extract meaningful information from even the simplest geometric representation of structures.

1.1. *Remote sensing of urban landscape*

To provide timely monitoring of cities and to classify urban land use over large areas, most research has utilised remotely sensed data from airborne or space-borne sensors. Projects such as the Global Human Settlement Layer (Pesaresi et al., 2013) and the Global Urban Footprint (Esch et al., 2013) have expanded this type of monitoring to global scales. Work has also expanded at more local spatial scales, attempting to classify, and monitor land uses within urban areas (Graessar et al., 2012; Kuffer & Barros, 2011) or to extract buildings and identify settled areas (Cheriyadat, 2014; Gros & Tiecke, 2016; Yuan, 2016). These local scale methods have been applied to monitoring informal housing in developing countries by measuring the distinctive patterns of small, dense, irregularly shaped, agglomerative structures (Graessar et al., 2012; Kit, Lüdeke, & Reckien, 2012; Kuffer, Pfeffer, & Sliuzas, 2016; Kuffer, Pfeiffer, Sliuzas, & Baud, 2016). Research in this area of intra-urban classification is notable for its methodological shift away from pixel-based spectral measures (e.g. vegetation indices) toward object-oriented feature extraction and spatial and textural measures that take advantage of patterns and textures detectable in very high spatial resolution (VHR) imagery (Herold et al., 2003; Tatem & Hay, 2004). The most commonly used textures in recent urban mapping applications include entropy, contrast, variance and other measures calculated on the grey level co-occurrence matrix (GLCM; Haralick, Shanmugam, & Dinstein, 1973) and related metrics (Pesaresi & Gerhardinger, 2011; Pesaresi, Gerhardinger, & Kayitakire, 2008) that delineate built-up areas (Duque et al., 2015; Kuffer, Pfeiffer, Sliuzas, & Baud, 2016; Owen & Wong, 2013). Other work has made use of lacunarity measures to quantify the spacing between structures (Kit et al., 2012) as well as the distribution and orientation of line segments extracted from the image (Engstrom et al., 2015). The complexity of urban settlement patterns often requires multiple metrics to be used together and different size filters or feature calculation windows to measure characteristics expressed at different spatial scales (Graessar et al., 2012). The growth in the availability of both VHR imagery and computing power needed to process it has made this area of research very active in recent years.

1.2. *Vector data analyses of urban areas*

Similar to the increasing availability of remote sensing datasets, geospatial vector data (i.e. points, lines, polygons) are now commonly collected and maintained for urban areas by government agencies as part of planning, topographic map production, and tax records, by commercial data providers, or even by volunteers as in the OpenStreetMap project (http://www.openstreetmap.org). These databases have varying degrees of completeness (Hecht, Kunze, & Hahmann, 2013), but, when they can provide comprehensive coverage of urban infrastructure, they offer an alternative approach from remote sensing imagery to monitor urban form and land uses.

In their richest and most complete form, vector databases can construct complete digital, 3D city models containing representations of individual structures. Such a model can add important information on building height to 2D representations on maps (Sridharan & Qiu, 2013). These databases can be time consuming and difficult to construct, however, and research has focused on building them through automated extraction from aerial photographs or LIDAR data (Rottensteiner & Briese, 2002). Other common vector data formats are 2D polygons delineating building footprints as commonly seen on topographic maps and cadastral surveys. While indicating the size and shape of a structure, these data rarely provide other information on land use or building height unless they can be linked with property data or tax records. In their most basic form, buildings can be represented as individual point features. Such settlement points (sometimes called dwelling unit points or address points) are most conventionally used to improve geocoding accuracy (Zandbergen, 2008), but they have also been used as ancillary datasets to identify settled areas for population distribution models (Zandbergen, 2011).

Yet characteristics of vector geometries can be indicative of land use in local areas. This idea requires an alternative interpretation of geospatial vector data – rather than representing discrete objects, the mapped shapes act as markers that, taken together as a pattern, identify broader or more general features of the built landscape. According to Steiniger et al. (2008), spatial pattern recognition of urban land uses adheres to principles of Gestalt psychology and human perceptions of form. When we view a topographic map, for example, we not only see individual structures, we also interpret patterns based on the proximity and similarity among objects to recognize concepts such as “suburbs” or “city centres.” Quantifying these patterns with building density, size, shape, and orientation can enable us to train more realistic, automated classifications (Steiniger et al., 2008). In developing such an interpretation of spatial data, Barr et al. (2004) distinguish between categories of “morphological properties” and “spatial relations” to organise shape measurements. The first category includes geometric attributes such as area (volume in 3D) or compactness of the shape. The latter group of spatial relationships or spatial structures includes measures of proximity or connectivity between vector objects which can be quantified with the number of edges and distances between nodes on a Gabriel graph or other spanning tree structure (Barr et al., 2004). This idea of pattern recognition and classification in spatial data has been taken up particularly by cartographers seeking to identify building types and to automate map generalisations (Hecht et al., 2015; Li, Yan, Ai, & Chen, 2004; Lüscher & Weibel, 2013; Steiniger et al., 2008; Zhang, Ai, Stoter, Kraak, & Molenaar, 2013).

In contrast to the studies discussed above, which all use 2D or 3D polygon representations of buildings, Longley and Mesev (2000) and Mesev (2005, 2007) demonstrated the use of point representations of structures for similar classification goals. Using address points of several UK cities from an Ordnance Survey database and point pattern statistics of density and nearest neighbour index, they identified measureable differences between UK neighbourhood types corresponding with construction years. The types of measures that can be calculated from point data are necessarily limited. Morphological properties such as compactness are not available for point geometries. Only spatial structures can be calculated and even then connectivity of the building structures (e.g. buildings sharing a wall) cannot be observed.

This current study emerges from the research stream of studies such as Mesev (2005, 2007) which use point-level vector data representing structures, yet its objectives are more closely aligned to those of remote sensing-based image classification of settlement areas, such as Graessar et al. (2012). Unlike previous vector data analyses (e.g. Barr et al., 2004; Hecht et al., 2015), the goal here is not to classify building features themselves into types, but to derive a surface classifying areas of particular settlement types. We begin with a land use map which covers portions of major cities in Afghanistan, yet we want to predict those basic categories for residential types in other areas. Section 2 develops a set of measures of spatial interrelationships between points which are calculated across scales and then used as data features in a machine learning method to classify settlement area types. The processing steps are computationally intensive and we discuss several steps to improve efficiency through parallelisation. We demonstrate our methods with data from seven provinces in Afghanistan.