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

Quantifying the geographic distribution of building coverage across the US for urban sustainability studies

A. Soliman[a,*], A. Mackay[b], A. Schmidt[c], B. Allan[b], S. Wang[a,d]

[a] National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, United States
[b] Department of Entomology, University of Illinois at Urbana-Champaign, United States
[c] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, United States
[d] Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, United States

ARTICLE INFO

Keywords:
Building coverage ratio
Green infrastructure
Urban intensification
NLME
Urban sustainability

ABSTRACT

The geographic distribution of building coverage is an important factor when evaluating the potential for green infrastructure and other sustainable development practices. However, comprehensive building coverage datasets are not available for many urban communities. We predicted the distribution of building coverage ratio across the contiguous United States (CONUS) at the scale of census block group using a bootstrap strategy applied to a nonlinear mixed effects model. Building footprints from 12 metropolitan areas were examined for completeness and aggregated at the level of U.S. census block groups ($n = 19{,}109$) and used to train the model. The best predictive model included both the percentage of the impervious surface area and housing unit density as nonlinear fixed effects in the form of a multivariate power function. Urban-type class, defined by median construction year of housing units and land use composition, were included in the model as a random effect. Cross-validation indicated that the selected model has a mean error of 0.049% (95% CI 0.047 and 0.051) of the estimated proportion of block group land area represented by building coverage. Adopting a bootstrapping strategy allowed selection of a subset with a minimum distance between samples to avoid spatial autocorrelation of residuals, while repeating the sampling for 1000 iterations to estimate the confidence intervals of the model parameters. Results are provided as open access data which quantify the geographical distribution of building coverage in terms of the ensemble predictions, mean and standard deviation. In addition, we describe trends in the model parameters across Urban-type classes. Although they were included as random effects in our predictive model, they will provide the urban research community with quantitative information about the effect of neighborhood age and type on the degree of urban intensification at a national scale.

1. Introduction

The fraction of the land area occupied by buildings excluding area represented by roads, parking lots, etc. (herein identified by the term “building coverage”) is a critical parameter for assessing the potential for green infrastructure for stormwater management (GI) (Aladenola & Adeboye, 2010) and other urban sustainable practices (e.g. solar energy, green corridors and urban heat island mitigation). For example, application of GI tools occupying a large land area and requiring a substantial setback from existing building structures (e.g. constructed wetlands) is generally limited to locales with a low building coverage, whereas GI tools with a smaller footprint, such as rain gardens, can be appropriate where building coverage is of a more moderate scale (Angel, Parent, Civco, & Blei, 2012). In locations with a very high building coverage, applications intimately associated with building structures (e.g. green roofs, rain cisterns) are likely to be the most suitable GI tools (Barron, Neis, & Zipf, 2014). Similarly, suitability assessments of other urban sustainable practices (e.g. solar energy (Bratsolis, Charou, Tsenoglou, & Vassilas, 2016), green corridors and urban heat island mitigation (Dietz & Clausen, 2008)) can be greatly informed by estimates of building coverage. However, the building coverage has not been characterized in many locations in the U.S., hindering attempts to evaluate the suitability of urban sustainability practices across regional or national scales. Although studies have shown that policymakers would benefit from national inventories of urban environments at local scales (Fan, Zipf, Fu, & Neis, 2014; Franco, Mandla, & Rao, 2016) and from inventories that capture the temporal dynamics of built-up areas (Fu, Zhang, Sharma, Pang, & Wang, 2017; Gagnon, Margolis, Melius, Phillips, & Elmore, 2016).

A review of the urban planning and GI literature reveals different

* Corresponding author at: National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, 1205 W. Clark St., MC-257, Urbana, IL 61801, United States
E-mail address: asoliman@illinois.edu (A. Soliman).

https://doi.org/10.1016/j.compenvurbsys.2018.05.010
Received 8 October 2017; Received in revised form 20 March 2018; Accepted 22 May 2018
0198-9715/

Please cite this article as: Soliman, A., Computers, Environment and Urban Systems (2018), https://doi.org/10.1016/j.compenvurbsys.2018.05.010

Table 1
Metrics of the geographic extension of man-made structures.

Nomenclature	Definition
Total Building Area	Total surface area occupied by building (Bratsolis, Charou, Tsenoglou, & Vassilas, 2016)
Available Roof Area	Total roof area that is suitable for a certain Green Infrastructure application (4; 10).
Building Coverage Ratio	The ratio of the total building area occupied = by buildings to total land area (19; 21; 20).
Floor Area Ratio	The ratio of gross building floor area to the total land area (19; 21; 20).
Housing Units Density	The number of housing units to the total land area (13; 24).
Percentage of Impervious Surface Area	The percentage of impervious surface area to the total land area (Haklay, 2010).

metrics used to describe the extent of man-made structures in a geographic region, along with conflicting definitions for these metrics (Table 1). Furthermore, data to define these metrics at the national level are not widely available (Groves, 2011). For example, the National Land Cover Database (NLCD) (Haklay, 2010) provides a percentage of impervious surface area at the continental US scale without differentiating between building structures and other built surfaces. Similarly, the Atlas of Urban Expansion is an open-source online resource that uses satellite data and detailed spatial analysis to derive metrics of built-up area in major cities around the world (Haklay & Weber, 2008), though it does not distinguish building area from the area represented by roads, parking lots, etc. A more recent dataset provided the density in terms of count of residential housing units per acre (Hamner, 2017). Open Street Maps (OSM) is an alternative data source which provides an outline of the perimeter of individual buildings and therefore could be used to estimate the building coverage (Hecht, Kunze, & Hahmann, 2013). However, the completeness and consistency of OSM are not fully quantified (Holm, 1979). A number of studies have evaluated OSM integrity in small regions, most notably in Germany using indirect measures like the number of contributing users and date of contributions (Homer et al., 2015; Iceland & Steinmetz, 2003; Izquierdo, Rodrigues, & Fueyo, 2008). Yet a complete assessment of OSM consistency in the United States has not been performed. In addition, there are methodological challenges to extracting building footprints directly from remote sensing data across a wide coverage area.

Different methods have been used in the past to estimate building coverage. Reported methods in the literature can be classified into direct and indirect methods. Direct methods rely heavily on remote sensing observations of buildings footprints from a variety of sensors (e.g. Optical, LiDAR, and Radar), while indirect methods rely on estimating the relationship between building coverage and other factors via regression. Izquierdo et al. (Bratsolis, Charou, Tsenoglou, & Vassilas, 2016) identified general requirements for developing a method to estimate building coverage and roof area as follows: (1) accurate; (2) estimate bounding error of the roof area estimates; (3) inexpensive; (4) efficient (low calculation times); (5) require few, global, available and standard input data; (6) produce georeferenced results; and (7) be scalable from local to global scales.

Despite the availability of remote sensing data, extracting, building footprint(BF), remains a difficult task. For example, the accuracy of outlining building height is severely affected by occlusions and shadows from man-made and natural objects in the urban areas (Fan, Zipf, Fu, & Neis, 2014; Khader & Montalto, 2009). In order to overcome this issue, Bratsolis et al. (Khader & Montalto, 2009) combined LiDAR with color orthophotos to overcome the problem of precision in estimating building coverage and height using LiDAR data alone. Similarly, an object-based classification was applied to LiDAR images to extract building coverage and building floor area ratio and identify buildings (Kuhn, 2008). Simple approaches were also reported to estimate the building coverage ratio such as digitizing the roof surface from high-resolution images and correcting for height displacement (Lehmann & Peter, 2003). However, the application of most direct methods is restricted to small areas because of the complexity of these methods.

Indirect methods estimate relative or absolute building area within a defined spatial unit (e.g. city block, neighborhood, town) using available confounding variables without the need to define the area of individual buildings. For example, Lehmann and Peter (Pan, Zhao, Chen, Liang, & Sun, 2008) estimated roof area of residential buildings across the EU using population density as a predictor. They developed a second order nonlinear polynomial model using microdata from the Northrine-Westfalia region in Germany. Franco et al. (Parshall et al., 2010) applied Support Vector Machine Classifier applied to Landsat 8 OLI images to identify the urban area in general and estimated the roof area by applying fixed coefficients to restrict the available urban area. While Izquierdo et al. (Bratsolis, Charou, Tsenoglou, & Vassilas, 2016) stratified building's typologies using population and building densities for obtaining a stratified random sample to estimate building coverage area using visual photo interpretation. The latter example combines both direct and indirect strategies and shows promise in estimating regional trends in building coverage.

Although there are efforts from numerous major metropolitan cities in the United States to make building coverage information accessible, there is not a complete national dataset available for public use. The first objective of this work was to develop a model for estimating building coverage in urban areas using public land use and demographic datasets available at the national scale. The outcome dataset will provide a descriptor of building coverage that is both uniformly available throughout the U.S. at the level of census block groups with a consistent definition and calculation. This dataset will be essential for efforts to assess the potential for urban sustainability applications at a national scale. A second objective was to explore the functional relationship between building coverage and other confounding variables such as population density, the density of housing units, road density and percentage of the impervious surface area and tree canopy. Describing these functional relationships is valuable in comparing the infrastructure of urban areas across the U.S.

2. Methods

2.1. Datasets and spatial data preparation

Multiple building footprint spatial datasets representing a broad range of community size and development intensity across the U.S. were obtained from public sources and pooled for twelve urban centers (Appendix A). Where BF dataset coverages overlapped, duplicate BF polygons were identified and eliminated. Block group (BG) polygons (R. C. Team, 2014) (n = 19,109) with at least half of their land area within a 2010 U.S. Census Urban Area (Reyna & Chester, 2015) and completely within our BF coverage area (confirmed using visual inspection of high-resolution satellite image; ESRI World Imagery) were then selected for developing the model. For each selected census block group, the sum area of BF polygons within the polygon boundary was calculated as a percentage of the total land area within a census block group. Land area was defined as the total area of block group area not identified in the 2011 National Land Cover Database (Haklay, 2010) as open water or wetland land cover at the spatial resolution of 30 m.

The spatial intersection was conducted by intersecting the individual buildings footprints with the block group boundaries. The intersection with the 2011 NLCD dataset was conducted by reclassifying land cover class to a binary value (land, water), and converting the reclassed raster to a vector file with polygons defined by the boundary of contiguous pixels with the same reclass value. Block group land area was then delineated by subtracting the area of overlap with polygons representing NLCD water.

Discrete land cover classes and continuous estimates of percentage