Modelling built-up expansion and densification with multinomial logistic regression, cellular automata and genetic algorithm

Ahmed Mustafa[a,b,*,1,2], Alison Heppenstall[c], Hichem Omrani[d], Ismaïl Saadi[a], Mario Cools[a], Jacques Teller[a]

[a] LEMA, Urban and Environmental Engineering Department, Liège University, Belgium
[b] Department of Computer Science, Purdue University, USA
[c] School of Geography, University of Leeds, UK
[d] Urban Development and Mobility Department, LISER, Luxembourg

ARTICLE INFO

ABSTRACT

This paper presents a model to simulate built-up expansion and densification based on a combination of a non-ordered multinomial logistic regression (MLR) and cellular automata (CA). The probability for built-up development is assessed based on (i) a set of built-up development causative factors and (ii) the land-use of neighboring cells. The model considers four built-up classes: non built-up, low-density, medium-density and high-density built-up. Unlike the most commonly used built-up/urban models which simulate built-up expansion, our approach considers expansion and the potential for densification within already built-up areas when their present density allows it. The model is built, calibrated, and validated for Wallonia region (Belgium) using cadastral data. Three 100 × 100 m raster-based built-up maps for 1990, 2000, and 2010 are developed to define one calibration interval (1990–2000) and one validation interval (2000 − 2010). The causative factors are calibrated using MLR whereas the CA neighboring effects are calibrated based on a multi-objective genetic algorithm. The calibrated model is applied to simulate the built-up pattern in 2010. The simulated map in 2010 is used to evaluate the model's performance against the actual 2010 map by means of fuzzy set theory. According to the findings, land-use policy, slope, and distance to roads are the most important determinants of the expansion process. The densification process is mainly driven by zoning, slope, distance to different roads and richness index. The results also show that the densification generally occurs where there are dense neighbors whereas areas with lower densities retain their densities over time.

1. Introduction

Built-up development is the most typical form of land-use change. Without policy interventions, built-up developments may cause destructive impacts on the environment, on natural resources and on human health (Zhang, Ban, Liu, & Hu, 2011). Consequently, modelling built-up development is attracting attention of scientists, urban planners and politicians alike. Most built-up/urban models (e.g. Han & Jia, 2017; Liao et al., 2014; Liu et al., 2014; Puertas, Henríquez, & Meza, 2014; Vermeiren, Van Rompaey, Loopmans, Serwajja, & Mukwaya, 2012) are raster-based with a coarse cell space ranging from 30 × 30 m to 300 × 300 m. Whilst many authors advocate a larger grid cell for land-use modelling, for example 100 × 100 m (e.g. Jiang, Liu, Yuan, & Zhang, 2007; Munshi, Zuidgeest, Brussel, & van Maarseveen, 2014; Poelmans & Van Rompaey, 2010), land-use cells with these dimensions usually comprise a mix of different land-uses (Omrani, Abdallah, Charif, & Longford, 2015). For example, a cell classified as built-up land may be occupied by 80% built-up surface and 20% arable surface. With increases in the spatial resolution of data, researchers have begun to use grid cells as small as 10 × 10 m, such as Berberoğlu, Akın, and Clarke (2016) model for Adana city (Turkey). However, the drawback to using such a fine resolution is that it requires intensive computational resources to model larger study areas such as regions where 100 × 100 m cell dimensions are commonly used (e.g. Omrani et al., 2015; Poelmans & Van Rompaey, 2010). One solution to address the trade-off between coarse regular cell spaces and heterogeneity is

---

* Corresponding author at: LEMA, Urban and Environmental Engineering Department, Liège University, Belgium.
E-mail addresses: a.mustafa@ulg.ac.be, a-mustafa@purdue.edu (A. Mustafa), a.j.heppenstall@leeds.ac.uk (A. Heppenstall), hichem.omrani@liser.lu (H. Omrani), ismail.saadi@ulg.ac.be (I. Saadi), mario.cools@ulg.ac.be (M. Cools), jacques.teller@ulg.ac.be (J. Teller).
[1] Permanent address: Liège University, Allée de la Découverte 9, Quartier Polytech 1, 4000 Liège, Belgium.
[2] Present address: Purdue University, Department of Computer Science, 305 N University St, West Lafayette, 47,907 IN, USA.

examining several built-up densities instead of a binary classification (i.e. non built-up/built-up).

Although built-up densification processes, transitions from low-density to high-density, is critically important for policy makers who are concerned with restricting sprawl (Nabielek, 2012; Tachieva, 2010), the literature on urban/built-up expansion models highlights that many of the models focus only on expansion process (e.g. Poelmans & Van Rompaey, 2009; Wang et al., 2013). However, there are a limited number of studies that consider the expansion of several urban densities and/or densification in a variety of ways. Mustafa, Cools, Saadi, and Teller (2015), Robinson, Murray-Rust, Rieser, Milicic, and Rounsevell (2012), Sunde, He, Zhou, Hubbart, and Spicci (2014), Xian and Crane (2005), Yang (2010) and Zhang et al. (2011) model the expansion of different urban/built-up densities. Crols et al. (2015), Loibl and Toetzer (2003), White, Engelen, and Uljee (2015) and White, Uljee, and Engelen (2012) model the processes of urban expansion as well as of densification. They define densification as an increase in population and/or several economic sectors density.

One of the most popular techniques of existing urban/built-up expansion models which are employed to analyze and/or predict the built-up pattern is cellular automata (CA) (e.g. Berberoğlu et al., 2016; Feng, Liu, Tong, Liu, & Deng, 2011; Han, Hayashi, Cao, & Imura, 2009; Tian et al., 2016; Wang et al., 2013). CA is a dynamic discrete space and time bottom-up modelling approach. CA is widely used in urbanization modelling due to its simplicity, transparency and powerful capacities for dynamic spatial simulation (Clarke & Gaydos, 1998). Aburas, Ho, Ramli, and Ash'aari (2016) and Santé, García, Miranda, and Crecente (2010) reviewed CA urbanization models concluding that the CA modelling approach is one of the most appropriate techniques for simulating urban/built-up patterns. However, key challenges in CA are calibrating the transition rules of built-up development probability as a function of (i) a series of causative factors (driving forces) and (ii) spatial (neighborhood) characteristics. Early methods for CA calibration are based on trial and error (e.g. White & Engelen, 1997) and/or a visual test, to determine the model's parameters (e.g. Clarke, Hoppen, & Gaydos, 1997; Ward, Murray, & Phinn, 2000). Recently, a variety of automated methods based on statistics (e.g. García, Santé, Boullón, & Crecente, 2013), machine learning (e.g. Rienow & Goetzke, 2015), artificial neural networks (e.g. Berberoğlu et al., 2016) and search algorithms for optimization such as genetic algorithms (e.g. Al-Ahmadi, See, Heppenstall, & Hogg, 2009) and particle swarm optimization (e.g. Feng et al., 2011) have begun to be widely employed.

Validation of CA models is another challenge. A common validation method is based on pixel-by-pixel location agreement (e.g. Poelmans & Van Rompaey, 2009). This approach cannot discriminate between "near-miss" and "far-miss" errors which limit its ability to detect spatial patterns (Mustafa, Saadi, Cools, & Teller, 2014). Another approach is based on spatial metrics (e.g. Roy Chowdhury & Maithani, 2014). Spatial metrics can be potentially misleading, for example, two areas with distinctly different infrastructures may show the same spatial index (White & Engelen, 2000). A third method is based on a fuzzy set theory. Fuzzy map comparison provides a method of dealing and comparing maps containing a complex mixture of spatial information (Ahmed, Ahmed, & Zhu, 2013). It takes into account local variations meaning that matches found at shorter distances are given a higher agreement. It measures the similarity of a cell in a value between 0 (fully-distinct) and 1 (fully-identical). Thus, it can easily distinguish areas of minor errors from areas of major errors. Van Vliet et al. (2016) present a comprehensive survey of calibration and validation practices in land use change modelling.

This study contributes to research efforts that model built-up expansion and densification processes. We model the built-up expansion (non-built-up to one of built-up density classes) and densification (lower built-up densities to higher ones). The model is based on a hybrid approach which integrates logistic regression and CA modelling

approaches. The model is applied to Wallonia (Belgium). Belgian cadastral data (CAD) are used to generate three built-up maps for the years 1990, 2000 and 2010. These maps represent four built-up classes: non built-up (class-0), low-density (class-1), medium-density (class-2) and high-density (class-3). Three maps can define one calibration interval (1990–2000) and one validation interval (2000–2010). The model considers a set of static causative factors related to accessibility, geo-physical features, policies and socio-economic factors. Another important factor is neighborhood interactions because of the fact that urbanization can be regarded as a self-organizing system (Poelmans & Van Rompaey, 2010).

The model's parameters are calibrated based on a logistic regression model and genetic algorithm. The logistic regression is employed to set the parameter of 12 built-up development causative factors: elevation, slope, zoning status, employment rate, richness index and Euclidian distances to highways, main roads, secondary roads, local roads, railway stations, large-sized and medium-sized Belgian cities. The richness index is calculated as the average income per capita for each municipality divided by the average income per capita in Belgium. The built-up causative factors are selected according to a literature survey of common factors involved in urban/built-up expansion models (e.g. Achmad, Hasyim, Dahlan, & Aulia, 2015; Cammerer, Thieken, & Verburg, 2013; Dubovyk, Sliuzas, & Flacke, 2011; Li, Zhou, & Ouyang, 2013; Poelmans & Van Rompaey, 2010; Verburg, van Eck, de Nijs, Dijst, & Schot, 2004) as well as the finding of previous studies conducted for Wallonia (Beckers et al., 2013; Mustafa et al., 2015). The dependent variable for the logistic regression model represents the changes from class-0 to class-1, class-2 or class-3, the changes from class-1 to class-2 and the changes from class-2 to class-3.

As the dependent variable is a multi-level, i.e. with more than two possible outcomes, we should consider a non-binary logistic regression. The most common logistic regression types that handle multiple levels of an outcome are ordered logistic regression and multinomial logistic regression. Ordered logistic regression assumes that the levels of dependent status have a natural ordering (i.e. low to high). This is known as the proportional odds model or parallel regression assumption (Kim, 2003). To evaluate this assumption, the test of the proportional odds assumption is performed. The null hypothesis of the test is that the relationship, i.e. coefficients, between each pair of dependent levels is the same. The significance of Chi-Square statistic of the proportional odds test is < 0.001. Given the assumption of having a natural ordering in the dependent variable is violated, thus a non-ordered multinomial logistic regression model (MLR) is adopted for this study.

A multi-objective genetic algorithm (MGA) is employed to calibrate the neighborhood interactions on a dynamic basis. García et al. (2013) reported that the GA is one of the most robust heuristic automated methods to solve optimization problems. A number of studies have used GA to calibrate CA models (e.g. Al-Ahmadi et al., 2009; García et al., 2013; Shan, Alkheder, & Wang, 2008). The MGA objective function is the maximization of allocation accuracy rates for all built-up classes. The accuracy rate function is defined as a fuzzy membership function of exponential decay with a halving distance of two cells and a neighborhood window of four cells. The accuracy rate function is also employed to validate the model.

2. Materials

2.1. Study area

The model is applied to Wallonia region, the southern part of Belgium (Fig. 1). Wallonia occupies an area of 16,844 km² and administratively consists of five provinces: Hainaut, Liège, Luxembourg, Namur, and Walloon Brabant. The total population in 2010 was 3,498,384 inhabitants, corresponding to one third of the Belgium population (Belgian Federal Government, 2013). The population is mainly