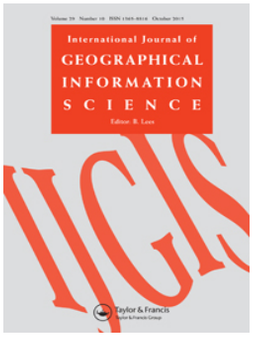

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

Simulating urban growth in a metropolitan area based on weighted urban flows by using web search engine

Jinyao Lin & Xia Li

To cite this article: Jinyao Lin & Xia Li (2015) Simulating urban growth in a metropolitan area based on weighted urban flows by using web search engine, International Journal of Geographical Information Science, 29:10, 1721-1736, DOI: 10.1080/13658816.2015.1034721

To link to this article: http://dx.doi.org/10.1080/13658816.2015.1034721

Published online: 23 Apr 2015.

International Journal of Geographical Information Science, 2015
Vol. 29, No. 10, 1721–1736, http://dx.doi.org/10.1080/13658816.2015.1034721

Simulating urban growth in a metropolitan area based on weighted urban flows by using web search engine

Jinyao Lin and Xia Li*

School of Geography and Planning, and Guangdong Key Laboratory for Urbanization and Geo-simulation, Sun Yat-sen University, Guangzhou, PR China

(*Received 19 October 2014; final version received 23 March 2015*)

As a consequence of rapid and immoderate urbanization, simulating urban growth in metropolitan areas effectively becomes a crucial and yet difficult task. Cellular automata (CA) model is an attractive tool for understanding complex geographical phenomena. Although intercity urban flows, the key factors in metropolitan development, have already been taken into consideration in CA models, there is still room for improvement because the influences of urban flows may not necessarily follow the distance decay relationship and may change over time. A feasible solution is to define the weights of intercity urban flows. Therefore, this study presents a novel method based on weighted urban flows ($CA_{WeightedFlow}$) with the support of web search engine. The relatedness measured by the co-occurrences of the cities' names (toponyms) on massive web pages can be deemed as the weights of intercity urban flows. After applying the weights, the gravitational field model is integrated with Logistic-CA to fulfill the modeling task. This method is employed to the urban growth simulation in the Pearl River Delta, one of the most urbanized metropolitan areas in China, from 2005 to 2008. The results indicate that our method outperforms traditional methods with respect to two measures of calibration goodness-of-fit. For example, $CA_{WeightedFlow}$ can yield the best value of 'figure of merit'. Moreover, the proposed method can be further used to explore various development possibilities by simply changing the weights.

Keywords: weighted urban flow; web search engine; gravitational field model; cellular automata; calibration

1. Introduction

Rapid urbanization has become an increasingly serious issue in developing countries, such as China (Cohen 2004, Grimm *et al.* 2008). Substantial economic development occurs in China at the expense of severe environmental and ecological problems since the adoption of the reform and opening-up policy in 1978, especially the country's three major metropolitan regions (i.e., the Pearl River Delta, the Yangtze River Delta, and the Beijing–Tianjin–Hebei region) (Liu *et al.* 2005, Tan *et al.* 2005). For example, the Pearl River Delta region contributed 9.2% of the gross domestic product (GDP) of the whole nation in 2012 (National Bureau of Statistics of China 2013), but has long been suffering from agricultural land loss (Yeh and Li 1999, Weng 2002), wetland loss (Li *et al.* 2006), environmental pollution (Fu *et al.* 2003, Zhang *et al.* 2011), etc. As a consequence, it is a crucial and meaningful task to better understand the spatio-temporal processes of urban growth in those large areas.

*Corresponding author. Email: lixia@mail.sysu.edu.cn

Cellular automata (CA) model is competent for the task and therefore has proliferated in modeling land-use and land-cover changes over the past two decades (Couclelis 1997, White *et al.* 1997, Batty *et al.* 1999, Verburg *et al.* 2004, He *et al.* 2006, Liu *et al.* 2008). Despite the fact that various CA models are still being continuously developed, most of which are suitable only for analyzing a single city (Li and Liu 2006). Such models are incapable of large-scale urban growth simulation because they mainly focus on the neighborhood functions (Santé *et al.* 2010). That is to say, they seldom examine distant influences among the cities within a metropolitan area. It is unreasonable to treat each city independently because metropolitan cooperation plays a growing role in regional development (Heeg *et al.* 2003, Luo and Shen 2009). To deal with this problem, intercity urban flows should be taken into account.

The term 'urban flow' is defined as the bidirectional or multidirectional interchanges of population, goods, information, capital, and technologies in a metropolitan area (Zhu and Yu 2002, Seto *et al.* 2012). Much effort has been made to analyze the influences of urban flows (e.g., Zhu and Yu 2002, Jiang *et al.* 2008). However, the application of urban flow factors in CA models is still rarely investigated. Until recently, He *et al.* (2013) tried to incorporate gravitational field model (GFM) with CA to simulate urban growth in the Beijing–Tianjin–Tangshan metropolitan area. The GFM was used to reflect the influence of intercity urban flows. They found that traditional CA models tend to overestimate the roles of neighborhood effect, and thus their proposed method can generate more accurate simulation results (He *et al.* 2013). Although this method considers urban flow factors in large-scale urban growth simulation, some drawbacks still remain unsolved. For example, the influences of urban flows from some cities on their neighboring cities may not necessarily follow the distance decay relationship (Fotheringham 1981, Tiefelsdorf 2003) and may change over time. It is because the interaction between the cities will differ due to various development policies and/or different time periods, etc. As inspired by the theory of weighted network (Batty 2013, Barrat *et al.* 2004, Newman 2004), a feasible solution is to define the weights of the influences of urban flows between every two cities.

However, quantifying proper weights for the relationships in complex city networks remains a tough challenge. Some researchers tried to estimate the relatedness between some cities based on big data, such as air passenger flow data (Xiao *et al.* 2013), mobile phone data (Phithakkitnukoon *et al.* 2010), and social media check-in data (Liu *et al.* 2014b). Unfortunately, those kinds of data are usually unavailable or costly for public use. Moreover, they also have their own limitations, social networking sites are more popular among young people for example (Correa *et al.* 2010).

Interestingly, a large body of literature indicates that geographical knowledge and information can be extracted from the World Wide Web, another important source of big data, with low cost (Buyukokkten *et al.* 1999, Jones *et al.* 2008, Shi and Barker 2011). Web search engine is such a tool that is designed to search for web information according to user-specified keywords. A notable example is Google Flu Trends, a web service to detect influenza epidemics based on search engine query data (Ginsberg *et al.* 2009). As a consequence, by using web search engine, Liu *et al.* (2014c) proposed an easy alternative to measure the relatedness between geographical entities from massive web pages. They simply recorded the numbers of web documents that contain both the place names (toponyms) of two entities and found that a high number of co-occurrences of toponyms in web documents indicate a strong relatedness between the places. In fact, the analysis of toponyms, which has long been applied in geographical information retrieval (Goodchild and Hill 2008), can discover geographical landscapes and processes (Luo *et al.* 2010). Obviously, the relatedness revealed by toponym

co-occurrences can be deemed as the weights of the influences of urban flows (i.e., weighted urban flows) between every two cities. Therefore, by combining previous works of He *et al.* (2013) and Liu *et al.* (2014c), this study proposes a novel method ($CA_{WeightedFlow}$ hereafter) to simulate urban growth in a metropolitan area based on weighted urban flows with the support of web search engine. What role can big data play in large-scale urban growth simulation? We will shed some light on this question in the era of big data.

2. Methodology

The urban outflows of the cities were first calculated based on the socioeconomic data. Second, we measured the relatedness between every two cities by using web search engine. A higher relatedness between the cities implies a higher weight of the influences of urban flows between them. The necessity for defining the weights is illustrated in Figure 1. After applying the weights, the influences of urban flows from the cities were reflected by the gravitational field model. Finally, the commonly used Logistic-CA model was employed to simulate urban growth in a metropolitan area based on those weighted urban flow factors. This study focuses on only the transition from non-urban to urban areas and assumes zero loss of urban. The succeeding subsections provide more details about the procedures.

2.1. *Urban flow*

Drawing on the concepts from Newton's law of universal gravitation (Verlinde 2011), the influence of urban flows from city i on the cell (x, y) can be approximately represented as follows (Liang 2009):

$$U_{i,x,y} = \frac{F_i}{D(x, y, x_i, y_i)} \tag{1}$$

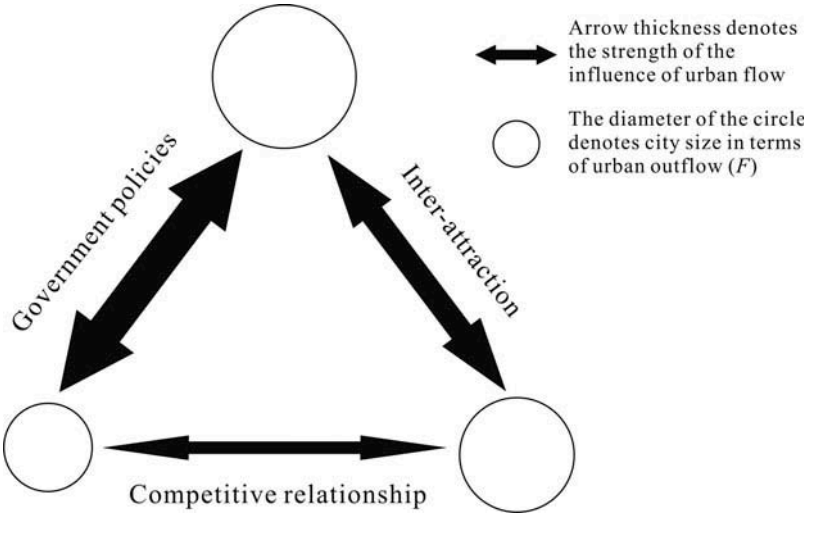

Figure 1. The necessity for defining the weights of intercity urban flows.

where $U_{i,x,y}$ is the urban flow intensity from city i to the cell (x, y), $D(x, y, x_i, y_i)$ is the Euclidean distance from city i's center (x_i, y_i) to the cell (x, y), and F_i is the urban outflow of city i (Zhu and Yu 2002, He *et al.* 2013):

$$F_i = N_i \cdot E_i \tag{2}$$

where N_i is the internal function for city i, which can be well represented by the city's GDP per employee; E_i is the external function for city i, which can be calculated using the equation given by:

$$E_i = \sum_{k=1}^{m} E_{ik} \tag{3}$$

where m is the number of outward-looking economic sectors, and E_{ik} is the output function of sector k of city i:

$$E_{ik} = G_{ik} - G_i \cdot \frac{G_k}{G} \tag{4}$$

where G_{ik} is the number of employees in sector k of city i, G_i is the number of employees in all the relevant sectors of city i, G_k is the number of employees in sector k of all the cities, and G is the number of employees in all the relevant sectors of all the cities. If $E_{ik} \leq 0$, it means that no output function can be derived from the economic sector k of city i, and E_{ik} will therefore be set to 0. Otherwise, the corresponding sector has an output function to support other cities (He *et al.* 2013).

He *et al.* (2013) suggested that urban flow factors for urban growth simulation can be estimated from Equation (1). However, we argue that it may be somewhat unreasonable to do so due to the complexity of city networks. To better quantify the urban flow intensity, the relatedness between the cities should be taken into consideration. The component of the relatedness for the urban flows is discussed in the following subsection.

2.2. *Relatedness between geographical entities*

The relatedness between two geographical entities can be easily measured by the co-occurrences of their names (toponyms) on massive web pages, specifically in news reports. The basic assumption lies behind this method is that a high number of co-occurrences of toponyms indicate a strong relatedness between the places (Liu *et al.* 2014c). After retrieving the co-occurrence data by using web search engine, the relatedness between every two cities (R_{ij}) can be standardized as follows:

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} \cdot C_{jj}}} \tag{5}$$

where C_{ij} is the number of co-occurrences of cities i and j, and C_{ii} (or C_{jj}) is the number of web pages retrieved using only one city's name i (or j).

The relatedness between the cities will be considered as the weights of the influences of urban flows. To be more exact, Equation (1) should be updated as follows:

$$U_{i,j,x,y} = \frac{F_i}{D_j(x, y, x_i, y_i)} \cdot R_{ij} \tag{6}$$

where $U_{i,j,x,y}$ is the weighted urban flow intensity from city i to the cell (x, y) within city j, $D_j(x, y, x_i, y_i)$ is the Euclidean distance from city i's center (x_i, y_i) to the cell (x, y) within city j, and F_i is the same as the one in Equation (1). It is also feasible if R_{ij} is measured by other big data-based methods mentioned in Section 1.

Subsequently, the urban flow factors generated through the above equation can be served as a major component for calibrating large-scale CA models. More detailed techniques will be presented in the following subsection.

2.3. *The framework of $CA_{weightedflow}$*

In this study, we only calibrated the CA models from two time points. The Logistic-CA model is selected to implement the simulation task because of its popularity and effectiveness (Wu 2002). Many improvements of the CA are based on this widely used model for simulating urban development and land-use changes. The methodology is the same if other types of CA are used. In Logistic-CA, the development potential of a non-urbanized cell is represented by using a logistic function (Li *et al.* 2013):

$$p_{ij}(S = \text{Developed}) = \frac{\exp(z_{ij})}{1 + \exp(z_{ij})} = \frac{1}{1 + \exp(-z_{ij})} \tag{7}$$

where p_{ij} is the potential of being in a developed state for cell ij, S is the state (developed or not), and z_{ij} is calculated based on a series of driving forces related to urban dynamics:

$$z_{ij} = a + \sum_k b_k S_k + \sum_m c_m U_m, \quad \text{for each cell } ij \tag{8}$$

where a is a constant, S_k is the kth spatial variable (e.g., distance to the city centers or road networks), U_m is the mth urban flow factor (generated through Equation (6)), and b_k, c_m are the corresponding parameters of S_k, U_m, respectively. These parameters can be obtained through logistic regression. The only difference between the common CA model and our proposed method is the selection of driving forces. That is to say, the former only considers proximity variables (i.e., $z_{ij} = a + \sum_k b_k S_k$), while the latter also includes intercity urban flow factors ($\sum_m c_m U_m$) simultaneously.

Moreover, the influences of stochastic factor, neighborhood effect, and geographical constraints should be incorporated in Equation (7). The new equation is given by:

$$p_{ij}^t = (1 + (-\ln \gamma)^\alpha) \times \frac{1}{1 + \exp(-z_{ij})} \times \Omega_{ij}^t \times con_{ij} \tag{9}$$

where p_{ij}^t is the probability of being in a developed state at time t for cell ij, γ is a stochastic factor ranging from 0 to 1, α is a parameter for controlling the stochastic degree, Ω_{ij}^t is the development density in a 3 × 3 Moore neighborhood of cell ij at time t, and con_{ij} is a constraint score for cell ij ranging from 0 to 1 (e.g., if a cell belongs to water areas, the score should be set to 0).

Subsequently, p_{ij}^t was compared with a threshold value to decide whether a non-urbanized cell should be developed to an urbanized cell during the iterations. The equation is given as follows:

$$S_{ij}^{t+1} = \begin{cases} \text{Developed}, & p_{ij}^t \geq p_{\text{threshold}} \\ \text{NonDeveloped}, & p_{ij}^t < p_{\text{threshold}} \end{cases} \quad (10)$$

where S_{ij}^{t+1} is the state of the cell ij at time $(t + 1)$, and $p_{\text{threshold}}$ is a threshold value determined by the total number of urbanized cells derived from the last observed (2008 in this study) Landsat Thematic Mapper (TM) images. However, the quantity of simulated change is unfixed because the number of converted cells is not bounded at each iteration. The iterations will stop once the quantity of simulated change exceeds the quantity of observed change.

Finally, we used the indicator of 'figure of merit' (FoM) to evaluate the performance of the simulation results. FoM can be calculated as follows (Pontius *et al.* 2008):

$$\text{Figure of merit} = \text{Hits} / (\text{Misses} + \text{Hits} + \text{False Alarms}) \times 100\% \quad (11)$$

where Misses are errors caused by observed urban gain simulated as non-urban persistence, Hits are agreements brought about by observed urban gain simulated as urban gain, and False Alarms are errors caused by observed non-urban persistence simulated as urban gain.

3. Implementation and results

3.1. Study area and spatial data

As shown in Figure 2, the Pearl River Delta (PRD) region is located in the central part of Guangdong Province, China. This fast-growing region includes nine administrative cities with a total area of approximately 55,000 km^2. The region has been experiencing rapid urbanization since the implementation of reform and opening-up policy in 1978. These tremendous land-use changes have given rise to a series of environmental and ecological problems (Yeh and Li 1999, Weng 2002, Fu *et al.* 2003, Zhang *et al.* 2011). Simulating urban growth in PRD is an urgent and meaningful endeavor.

After being rescaled to a spatial resolution of 150 m, the classified Landsat TM images in 2005 and 2008 were used to calibrate the CA model in this study. In addition, as presented in Figure 3, several spatial variables related to urban dynamics (S in Equation (8)) were also necessary for the calibration. These variables included the distances to the railways, highways, roads, city centers, and town centers, all of which were normalized into the range [0, 1]. The year of them is 2008, the end date of the calibration time period (Conway and Wellen 2011).

3.2. Implementation

First, the co-occurrence data were collected manually by using Baidu news search engine (http://news.baidu.com/) on 28 May 2014. Our search was restricted to the news from SINA Corporation (http://www.sina.com.cn), a leading media website serving China and the global Chinese communities, in accordance with related study (Liu *et al.* 2014c). We recorded the numbers of web pages that contain both the toponyms of each pair (every

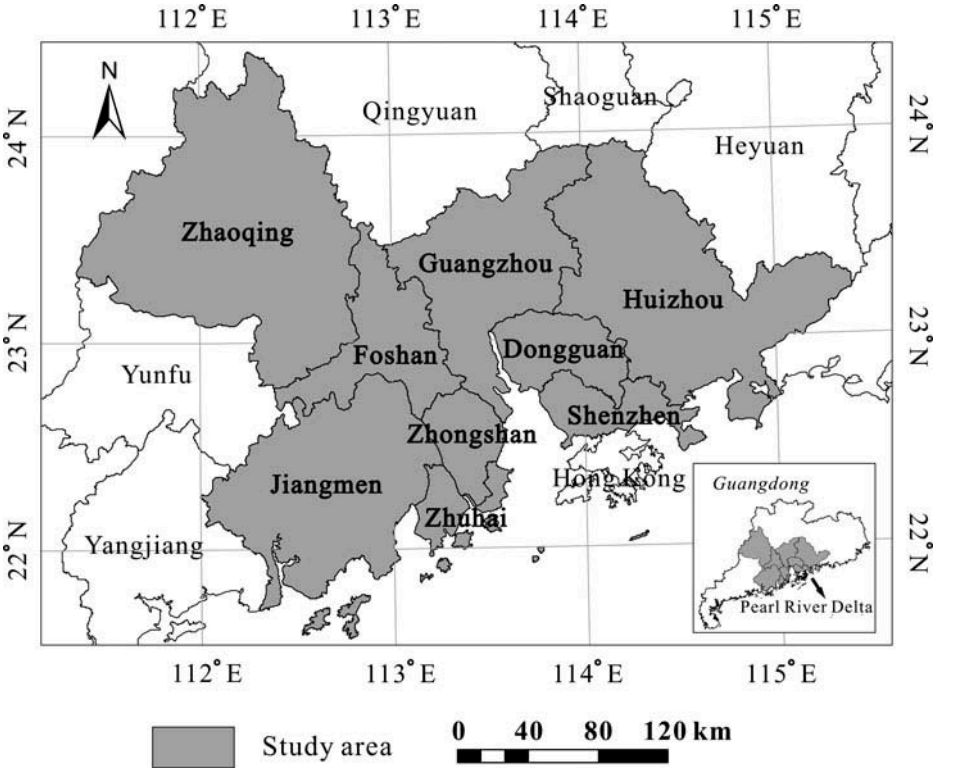

Figure 2. Location of the study area.

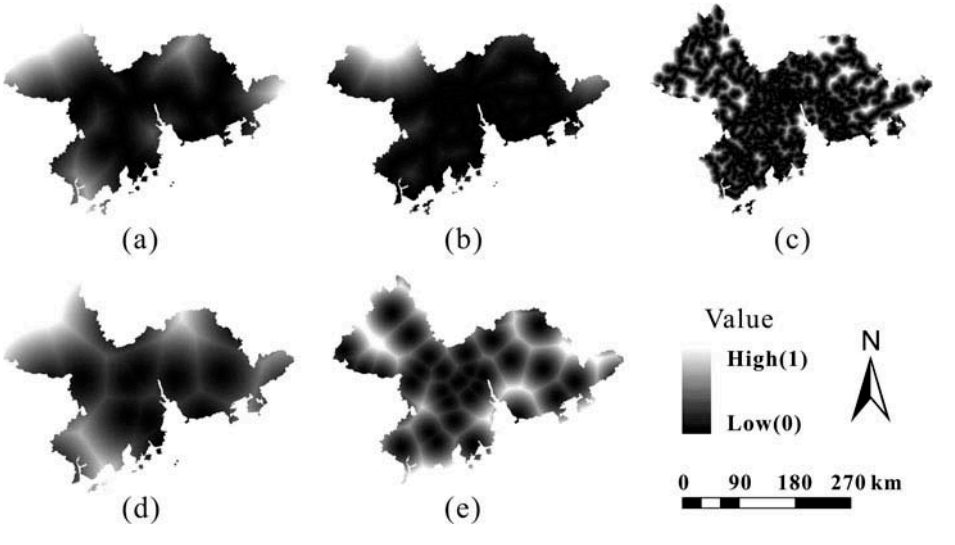

Figure 3. Various spatial variables related to urban dynamics in PRD: (a) distance to the railways, (b) distance to the highways, (c) distance to the roads, (d) distance to the city centers, and (e) distance to the town centers.

Table 1. Toponym co-occurrences and relatedness between every two cities in PRD.

	GZ	SZ	ZH	DG	FS	ZS	HZ	JM	ZQ
GZ	4490	168	6.03	47.5	92.5	62.9	2.99	1.63	1.92
SZ	0.0405	3840	6.37	67	3.95	2.56	5.88	0.47	0.64
ZH	0.0020	0.0023	2040	1.78	3.74	6.63	2.68	3.47	1.33
DG	0.0207	0.0316	0.0012	1170	3.55	2.42	5.09	0.54	1.09
FS	0.0361	0.0017	0.0022	0.0027	1460	3.64	2.21	4.14	1.98
ZS	0.0250	0.0011	0.0039	0.0019	0.0025	1410	1.13	2.62	0.45
HZ	0.0015	0.0032	0.0020	0.0051	0.0020	0.0010	866	3.33	2.71
JM	0.0011	0.0003	0.0034	0.0007	0.0047	0.0030	0.0049	524	3.31
ZQ	0.0014	0.0005	0.0015	0.0016	0.0026	0.0006	0.0045	0.0071	411

Notes: The values in the upper triangular matrix are the numbers of toponym co-occurrences in hundreds; the values in the diagonal entries are the numbers of web pages retrieved using only one city's name in hundreds; and the values in the lower triangular matrix are the relatedness of each pair.
Abbreviations (the same below): GZ – Guangzhou, SZ – Shenzhen, ZH – Zhuhai, DG – Dongguan, FS – Foshan, ZS – Zhongshan, HZ – Huizhou, JM – Jiangmen, ZQ – Zhaoqing.

two cities) from 2005 to 2008. Subsequently, the relatedness (weights) between the cities can be calculated according to Equation (5). The results are listed in Table 1.

Second, all the required statistical data were obtained from the Statistics Bureau of Guangdong Province. The numbers of employees from eight different outward-looking economic sectors related to intercity linkages were used to calculate the urban outflows (F in Equation (6)) of all the cities (Zhu and Yu 2002). These sectors are (1) Transport, storage, and postal services; (2) Wholesale and retail; (3) Finance; (4) Real estate; (5) Health care and social welfare; (6) Education and culture; (7) Scientific research and technical services; and (8) Manufacture. The calculation results are shown in Table 2.

After multiplying by the weights, the urban flow intensity from each city can be calculated according to Equation (6). The nine cities in PRD were classified into four levels according to their socioeconomic characteristics (i.e., GDP and population) (He *et al.* 2013): (1) Guangzhou, (2) Shenzhen, (3) Dongguan and Foshan, (4) Zhuhai, Zhongshan, Huizhou, Jiangmen, and Zhaoqing. When a cell is influenced by the urban flows from different cities within the same level, the maximum value is accounted (Wang *et al.* 2011). The final flow values from the four levels were then normalized into the range [0, 1]. These urban flow factors were deemed as the additional driving forces (U in Equation (8)). In addition, we also generated the original urban flow factors (without the weights) proposed by He *et al.* (2013), which were calculated based on Equation (1), for comparison. The results are displayed in Figure 4. It is found that our results are more reasonable than the traditional ones. For example, Foshan and Dongguan are on the west and east of Guangzhou, respectively. However, Guangzhou should exert more influence on Foshan than Dongguan because of the 'Guangzhou-Foshan City Integration' strategy. This phenomenon is consistent with the numbers of toponym co-occurrences in web documents.

Table 2. The urban outflows (F) of the cities.

	GZ	SZ	ZH	DG	FS	ZS	HZ	JM	ZQ
F	562.87	366.03	122.62	518.59	239.80	101.28	182.42	79.94	73.70

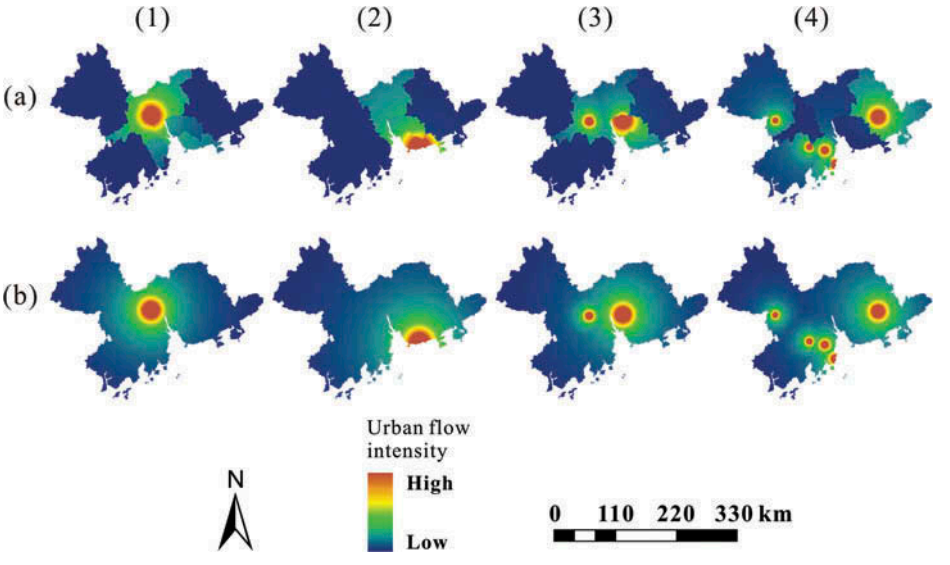

Figure 4. The urban flow intensity (U) from (1) Guangzhou, (2) Shenzhen, (3) Dongguan and Foshan, (4) Zhuhai, Zhongshan, Huizhou, Jiangmen, and Zhaoqing: (a) weighted urban flows and (b) unweighted urban flows.

Finally, we also constructed the model proposed by He *et al.* (2013) (CA_{Flow} hereafter), and the traditional Logistic-CA model ($CA_{Logistic}$ hereafter) for comparison. The latter one does not take urban flows into consideration. Both 1000 samples of urban land and 1000 samples of non-urban land were randomly collected from the observed Landsat TM images. All the CA models in this study were calibrated by using the same set of training samples for consistency. Next, logistic regression was used to obtain the parameters (a, b_k and c_m in Equation (8)) of the input driving forces. The results of the three methods are listed in Table 3. The development probability of the non-urbanized cells can then be calculated according to Equation (9). In order to reduce uncertainties brought

Table 3. The parameters of $CA_{WeightedFlow}$, CA_{Flow}, and $CA_{Logistic}$.

$CA_{WeightedFlow}$	a	$b_{TownCenter}$	$b_{CityCenter}$	$b_{Highway}$	$b_{Railway}$
	1.578	−2.238	−3.988	−2.983	0.492
	b_{Road}	c_{Flow1}	c_{Flow2}	c_{Flow3}	c_{Flow4}
	−8.358	28.375	358.416	368.223	212.893
CA_{Flow}	a	$b_{TownCenter}$	$b_{CityCenter}$	$b_{Highway}$	$b_{Railway}$
	0.561	−2.038	−3.199	−2.135	0.327
	b_{Road}	c_{Flow1}	c_{Flow2}	c_{Flow3}	c_{Flow4}
	−8.553	94.542	576.091	292.055	182.168

$CA_{Logistic}$	a	$b_{TownCenter}$	$b_{CityCenter}$	$b_{Highway}$	$b_{Railway}$	b_{Road}
	2.993	0.158	−7.025	−5.158	−0.504	−8.760

Notes: a is a constant, $c_{Flow1–4}$ denote the respective urban flow factors derived from city level (1)–(4), and the rest denote the corresponding proximity variables.

about by the stochastic factor, we repeated all the CA models 10 times. The average of the 10 repeated runs was considered as the final result, which will be assessed by comparing with the observed map of urban growth during 2005–2008 (Liu *et al.* 2014a).

3.3. Results

The performance of the three methods ($CA_{WeightedFlow}$, CA_{Flow}, and $CA_{Logistic}$) will be examined by two measures of calibration goodness-of-fit (Pontius and Pacheco 2004): cell-based accuracy (i.e., FoM) and the city's proportion of urban change to the total. The former is calculated through Equation (11), while the latter is measured as follows:

$$P_i = \frac{Q_i}{Q} \times 100\% \tag{12}$$

where P_i denotes city i's proportion of urban change to the total, Q_i is the quantity (area) of observed (or simulated) urban change of city i, and Q is the quantity of total observed (or simulated) urban change in the whole study area. Note that Q may slightly differ in different simulation results. We computed P_i for the reference maps and for each of the three model outputs, respectively.

Figure 5 displays the urban gain during 2005–2008 by focusing on a fast growing (Dongguan) and a slowly growing city (Jiangmen), and Figure 6 presents the visual comparison among the three simulation results. In general, the major fast growing cities account for most of the urban growth in a metropolitan area. However, the newly urbanized cells generated by the traditional Logistic-CA model were almost evenly distributed around preexisting urban patches (e.g., Figure 5c1 and c2). Instead, most of the newly urbanized cells of $CA_{WeightedFlow}$ and CA_{Flow} were distributed around several big cities (e.g., Figure 5a1 and a2). Subsequently, the indicator of FoM was used to quantify their performance. A box and whisker plot of FoM values among the 10 runs for each model is shown in Figure 7, while the FoMs of the average results are displayed in

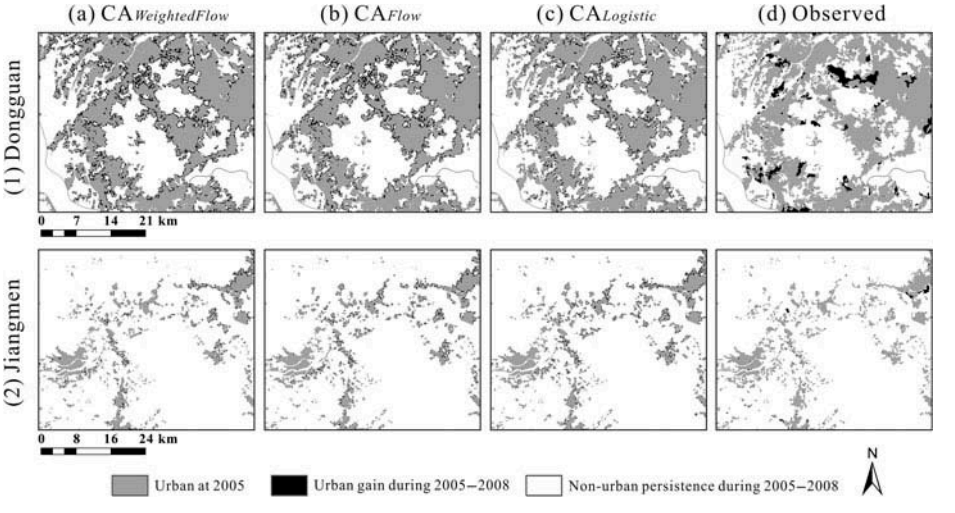

Figure 5. Urban change derived from the three simulation results and reference maps during 2005–2008: (1) Dongguan and (2) Jiangmen.

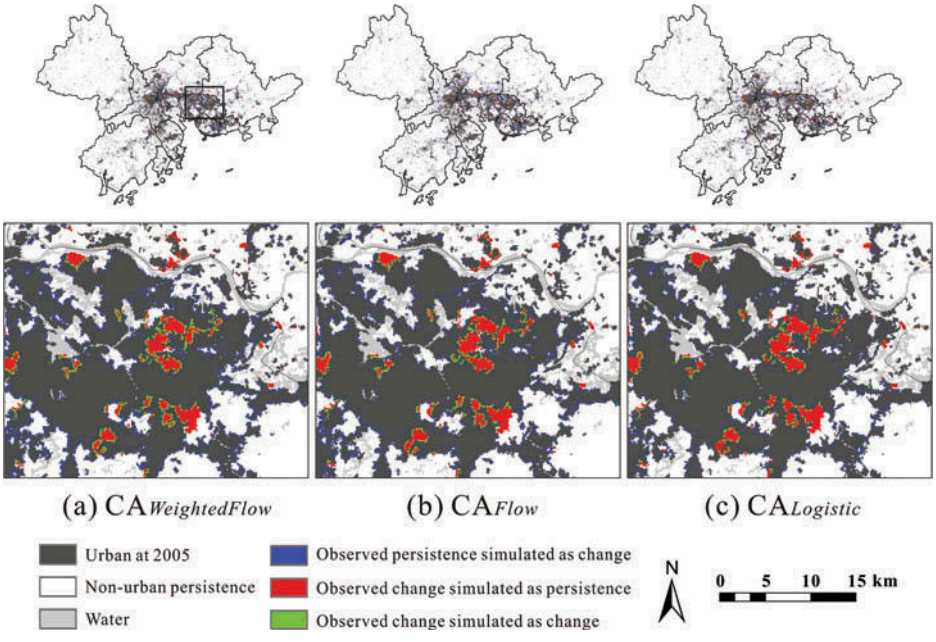

Figure 6. Visual comparison among the three simulation results of PRD: (a) $CA_{WeightedFlow}$, (b) CA_{Flow}, and (c) $CA_{Logistic}$.

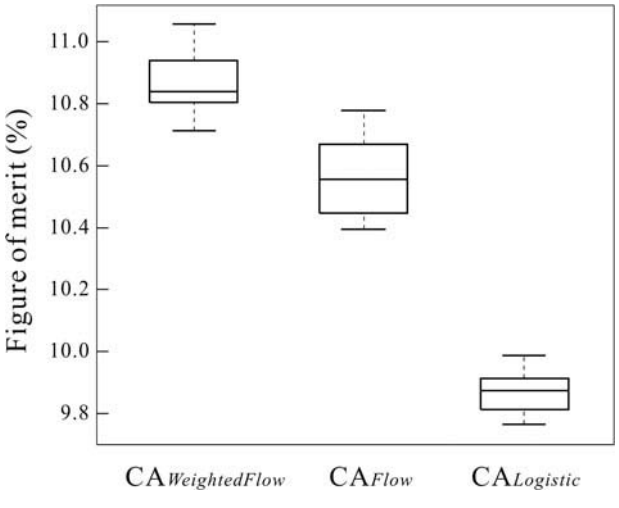

Figure 7. Box and whisker plot of FoM values among the 10 runs for each model.

Figure 8 (the FoM values in Figure 8 are greater than the individual ones in Figure 7 because the value of False Alarms in Equation (11) (partially caused by the stochastic factor) will decrease after averaging). It is found that $CA_{WeightedFlow}$ outperforms the other two.

We further focused on the top three fast-growing cities in PRD, namely Dongguan, Guangzhou, and Foshan, since these cities are more likely to be affected by severe

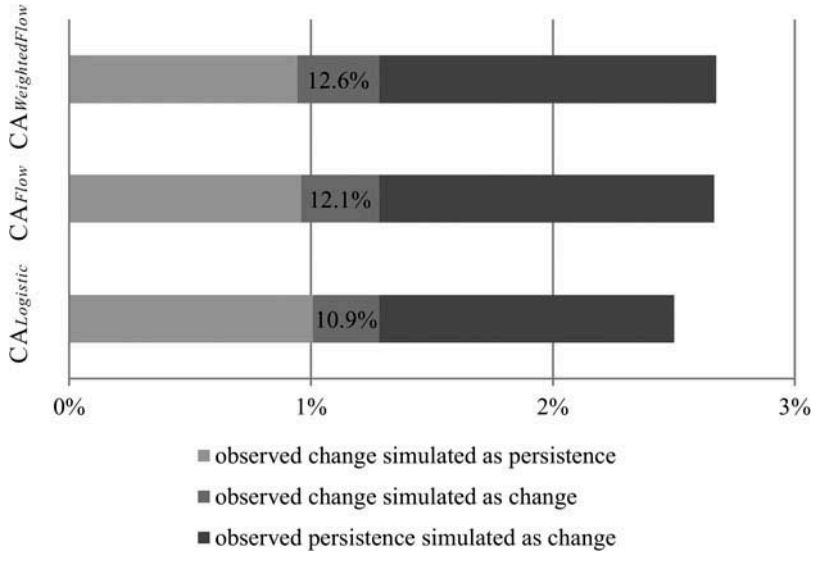

Figure 8. FoM derived from the average simulation result for each model. The unit of the horizontal axis is the percent of non-urban land at 2005.

Table 4. The proportion of urban change to the total (P_i) of the top three fast-growing cities in PRD during 2005–2008.

City Model	DG	GZ	FS	DG + GZ + FS	Average error
Observed	26.61% (1)	19.31% (2)	16.92% (3)	62.84%	0.00%
$CA_{WeightedFlow}$	23.93% (1)	22.36% (2)	14.80% (3)	61.09%	10.30%
CA_{Flow}	22.60% (2)	24.55% (1)	15.11% (3)	62.26%	13.46%
$CA_{Logistic}$	20.24% (2)	21.30% (1)	14.57% (3)	56.11%	14.71%

Note: The number in brackets denotes the city's growth ranking.

environmental and ecological problems. The total quantity of urban gain in PRD during 2005–2008 is 25,267 cells (equivalent to 568.51 km^2). As summarized in Table 4, these three cities account for 62.84% of the total growth. While $CA_{Logistic}$ is still the worst performing method, CA_{Flow} can yield a much closer value (62.26%). However, its respective proportions are quite different from those of the observed images. By contrast, $CA_{WeightedFlow}$ can generate more similar results than the other two models in terms of the average error.

$$\text{Average error} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\text{Simulated}_i - \text{Observed}_i}{\text{Observed}_i}\right| \times 100\% \tag{13}$$

where n is the number of attributes, Simulated_i denotes attribute i's value calculated from the model output, and Observed_i denotes attribute i's value calculated from the observed images.

In addition, Dongguan's and Guangzhou's P_i ranked first and second among the cities in PRD, respectively. But this observation can only be well reflected by $CA_{WeightedFlow}$.

3.4. Discussion

Based on the above experiments, we can conclude that the proposed method ($CA_{WeightedFlow}$) can yield the best simulation results, despite the fact that it only shows slight improvement over CA_{Flow} with respect to FoM. But more importantly, $CA_{WeightedFlow}$ can better characterize the complex processes of urban growth in major fast-growing cities in terms of the cities' proportions of urban change to the total. It is because the urban flow intensity for each city can be quantified more reasonably when taking into account the relatedness between the cities. Various big data-based methods have been developed to quantify relatedness between different regions, the calibration of large-scale urban CA models could benefit from those achievements.

In addition, the web search results can be restricted to a certain time period to better reflect the complex relationships at different stages of development. Finally, scenario analysis can be further conducted to predict future urban growth. For example, if the interaction between any two cities is to be strengthened or weakened, we can see what will happen by simply increasing or decreasing the corresponding weights. The results might contribute to urban planning in large areas. It may be difficult to explore these various possibilities without considering the weights between the cities.

4. Conclusions

Although intercity urban flows play an increasingly important role in urban growth in metropolitan areas, most of the CA models developed so far rarely take them into account. Recently, He *et al.* (2013) attempted to include urban flow factors by integrating gravitational field model with CA. However, the influences of urban flows may not strictly adhere to the distance decay relationship and may change over time. As a solution, this study has demonstrated that the aforementioned problems can be alleviated by defining the weights of the influences of intercity urban flows with the support of web search engine. We can easily measure the relatedness between the cities by a count of toponym co-occurrences in massive web documents. A higher number of co-occurrences indicates a higher relatedness (weight) and, consequently, stronger interaction between these two cities. The gravitational field model was then used to reflect the influence of intercity urban flows. Based on the co-occurrence data, our novel method can generate more reasonable urban flow factors than traditional method. These urban flow factors were deemed as the additional driving forces used in CA models (Table 3). Other procedures were the same as those of the common Logistic-CA models.

The proposed method ($CA_{WeightedFlow}$) was used to simulate urban growth in the Pearl River Delta region, one of the most urbanized metropolitan areas in southern China, from 2005 to 2008. The comparisons indicate that the results are slightly better than those obtained from two traditional CA models (i.e., CA_{Flow} and $CA_{Logistic}$), in terms of two evaluation metrics, namely cell-based accuracy (Figures 7–8) and the city's proportion of urban change to the total (Table 4). $CA_{WeightedFlow}$ can yield the best value of FoM and can better characterize the complex processes of urban growth in major fast-growing cities. Compared with traditional methods, our method may be more suitable and reasonable for calibrating large-scale urban CA models, in which big data play a key role. Nevertheless, further studies are still under way to overcome its remaining drawbacks. For example, it

may be unsuitable to apply the same set of transition rules for all the cities owing to heterogeneous geographical features. We will take knowledge transfer techniques into consideration.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant No. 41371376). We thank the editor and three anonymous reviewers for their useful comments and suggestions that greatly improved this paper.

References

Barrat, A., *et al*., 2004. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101 (11), 3747–3752. doi:10.1073/pnas.0400087101

Batty, M., 2013. *The new science of cities*. Cambridge, MA: MIT Press.

Batty, M., Xie, Y., and Sun, Z., 1999. Modeling urban dynamics through GIS-based cellular automata. *Computers, Environment and Urban Systems*, 23 (3), 205–233. doi:10.1016/S0198-9715(99)00015-0

Buyukokkten, O., *et al*., 1999. Exploiting geographical location information of web pages. *In*: *ACM SIGMOD workshop on the web and databases (WebDB'99)*, Philadelphia, PA. Stanford, CA: Stanford InfoLab, 1–6.

Cohen, B., 2004. Urban growth in developing countries: a review of current trends and a caution regarding existing forecasts. *World Development*, 32 (1), 23–51. doi:10.1016/j.worlddev.2003.04.008

Conway, T.M. and Wellen, C.C., 2011. Not developed yet? Alternative ways to include locations without changes in land use change models. *International Journal of Geographical Information Science*, 25 (10), 1613–1631. doi:10.1080/13658816.2010.534738

Correa, T., Hinsley, A.W., and Gil de Zúñiga, H., 2010. Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, 26 (2), 247–253. doi:10.1016/j.chb.2009.09.003

Couclelis, H., 1997. From cellular automata to urban models: new principles for model development and implementation. *Environment and Planning B: Planning and Design*, 24, 165–174. doi:10.1068/b240165

Fotheringham, A.S., 1981. Spatial structure and distance-decay parameters. *Annals of the Association of American Geographers*, 71 (3), 425–436.

Fu, J., *et al*., 2003. Persistent organic pollutants in environment of the Pearl River Delta, China: an overview. *Chemosphere*, 52 (9), 1411–1422. doi:10.1016/S0045-6535(03)00477-6

Ginsberg, J., *et al*., 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012–1014. doi:10.1038/nature07634

Goodchild, M.F. and Hill, L.L., 2008. Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22 (10), 1039–1044. doi:10.1080/13658810701850497

Grimm, N.B., *et al*., 2008. Global change and the ecology of cities. *Science*, 319, 756–760. doi:10.1126/science.1150195

He, C., *et al*., 2006. Modeling urban expansion scenarios by coupling cellular automata model and system dynamic model in Beijing, China. *Applied Geography*, 26 (3–4), 323–345. doi:10.1016/j.apgeog.2006.09.006

He, C., *et al*., 2013. Modeling the urban landscape dynamics in a megalopolitan cluster area by incorporating a gravitational field model with cellular automata. *Landscape and Urban Planning*, 113, 78–89. doi:10.1016/j.landurbplan.2013.01.004

Heeg, S., Klagge, B., and Ossenbruügge, J., 2003. Metropolitan cooperation in Europe: theoretical issues and perspectives for urban networking 1. *European Planning Studies*, 11 (2), 139–153. doi:10.1080/0965431032000072846

Jiang, B., Xiu, C., and Chen, C., 2008. Analysis of urban flow and model explanation on the urban group in middle-south part of Liaoning Province. *Economic Geography*, 28 (5), 853–861.

Jones, R., *et al*., 2008. Geographic intention and modification in web search. *International Journal of Geographical Information Science*, 22 (3), 229–246. doi:10.1080/13658810701626186

Li, X., *et al*., 2006. Inventory of mangrove wetlands in the Pearl River Estuary of China using remote sensing. *Journal of Geographical Sciences*, 16 (2), 155–164. doi:10.1007/s11442-006-0203-2

Li, X., *et al*., 2013. Calibrating cellular automata based on landscape metrics by using genetic algorithms. *International Journal of Geographical Information Science*, 27 (3), 594–613. doi:10.1080/13658816.2012.698391

Li, X. and Liu, X., 2006. An extended cellular automaton using case-based reasoning for simulating urban development in a large complex region. *International Journal of Geographical Information Science*, 20 (10), 1109–1136. doi:10.1080/13658810600816870

Liang, S., 2009. Research on the urban influence domains in China. *International Journal of Geographical Information Science*, 23 (12), 1527–1539. doi:10.1080/13658810802363614

Liu, J., Zhan, J., and Deng, X., 2005. Spatio-temporal patterns and driving forces of urban land expansion in China during the economic reform era. *AMBIO: A Journal of the Human Environment*, 34 (6), 450–455.

Liu, X., *et al*., 2008. A bottom-up approach to discover transition rules of cellular automata using ant intelligence. *International Journal of Geographical Information Science*, 22 (11–12), 1247–1269. doi:10.1080/13658810701757510

Liu, Y., *et al*., 2014b. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PloS One*, 9 (1), e86026.

Liu, Y., *et al*., 2014c. Analyzing relatedness by toponym co-occurrences on web pages. *Transactions in GIS*, 18 (1), 89–107. doi:10.1111/tgis.12023

Liu, Y., Feng, Y., and Pontius, R.G., 2014a. Spatially-explicit simulation of urban growth through self-adaptive genetic algorithm and cellular automata modelling. *Land*, 3 (3), 719–738. doi:10.3390/land3030719

Luo, W., Hartmann, J.F., and Wang, F., 2010. Terrain characteristics and Tai toponyms: a GIS analysis of Muang, Chiang and Viang. *GeoJournal*, 75 (1), 93–104. doi:10.1007/s10708-009-9291-8

Luo, X. and Shen, J., 2009. A study on inter-city cooperation in the Yangtze river delta region, China. *Habitat International*, 33 (1), 52–62. doi:10.1016/j.habitatint.2008.04.002

National Bureau of Statistics of China, 2013. *China statistical yearbook 2013*. Beijing: China Statistics Press.

Newman, M.E., 2004. Analysis of weighted networks. *Physical Review E*, 70 (5), 056131. doi:10.1103/PhysRevE.70.056131

Phithakkitnukoon, S., *et al*., 2010. Activity-aware map: identifying human daily activity pattern using mobile phone data. *In*: A. Ali Salah, *et al*., eds. *Human behavior understanding*. Berlin: Springer, 14–25.

Pontius, R.G., *et al*., 2008. Comparing the input, output, and validation maps for several models of land change. *The Annals of Regional Science*, 42 (1), 11–37. doi:10.1007/s00168-007-0138-2

Pontius, R.G. and Pacheco, P., 2004. Calibration and validation of a model of forest disturbance in the Western Ghats, India 1920–1990. *GeoJournal*, 61 (4), 325–334. doi:10.1007/s10708-004-5049-5

Santé, I., *et al*., 2010. Cellular automata models for the simulation of real-world urban processes: a review and analysis. *Landscape and Urban Planning*, 96 (2), 108–122. doi:10.1016/j.landurbplan.2010.03.001

Seto, K.C., *et al*., 2012. Urban land teleconnections and sustainability. *Proceedings of the National Academy of Sciences*, 109 (20), 7687–7692. doi:10.1073/pnas.1117622109

Shi, G. and Barker, K., 2011. Extraction of geospatial information on the Web for GIS applications. *In*: *10th IEEE international conference on cognitive informatics & cognitive computing (ICCI*CC)*, Banff, AB. IEEE: Piscataway, NJ, 41–48.

Tan, M., Li, X., and Lu, C., 2005. Urban land expansion and arable land loss of the major cities in China in the 1990s. *Science in China Series D: Earth Sciences*, 48 (9), 1492–1500.

Tiefelsdorf, M., 2003. Misspecifications in interaction model distance decay relations: a spatial structure effect. *Journal of Geographical Systems*, 5 (1), 25–50. doi:10.1007/s101090300102

Verburg, P.H., *et al*., 2004. Land use change modelling: current practice and research priorities. *GeoJournal*, 61 (4), 309–324. doi:10.1007/s10708-004-4946-y

Verlinde, E., 2011. On the origin of gravity and the laws of Newton. *Journal of High Energy Physics*, 2011 (4), 1–27. doi:10.1007/JHEP04(2011)029

Wang, L., *et al*., 2011. Research on urban spheres of influence based on improved field model in central China. *Journal of Geographical Sciences*, 21 (3), 489–502. doi:10.1007/s11442-011-0859-0

Weng, Q., 2002. Land use change analysis in the Zhujiang Delta of China using satellite remote sensing, GIS and stochastic modelling. *Journal of Environmental Management*, 64 (3), 273–284. doi:10.1006/jema.2001.0509

White, R., Engelen, G., and Uljee, I., 1997. The use of constrained cellular automata for high-resolution modelling of urban land-use dynamics. *Environment and Planning B: Planning and Design*, 24, 323–343. doi:10.1068/b240323

Wu, F., 2002. Calibration of stochastic cellular automata: the application to rural-urban land conversions. *International Journal of Geographical Information Science*, 16 (8), 795–818. doi:10.1080/13658810210157769

Xiao, Y., *et al*., 2013. Reconstructing gravitational attractions of major cities in China from air passenger flow data, 2001–2008: a particle swarm optimization approach. *The Professional Geographer*, 65 (2), 265–282. doi:10.1080/00330124.2012.679445

Yeh, A.G.O. and Li, X., 1999. Economic development and agricultural land loss in the Pearl River Delta, China. *Habitat International*, 23 (3), 373–390. doi:10.1016/S0197-3975(99)00013-2

Zhang, Y., *et al*., 2011. Procuring the regional urbanization and industrialization effect on ozone pollution in Pearl River Delta of Guangdong, China. *Atmospheric Environment*, 45 (28), 4898–4906. doi:10.1016/j.atmosenv.2011.06.013

Zhu, Y. and Yu, N., 2002. Urban flows in the HuNingHang urban compact district. *Urban Planning Forum*, 25 (1), 31–33.