SPATIAL DEPENDENCY OF URBAN SPRAWL AND THE UNDERLYING ROAD NETWORK STRUCTURE

Xianyuan Zhan[1] and Satish V. Ukkusuri[2]

[1]JD Intelligent City Research, JD Digits. Email: zhanxianyuan@jd.com. Formerly at Lyles School of Civil Engineering, Purdue University

[2]Professor, Lyles School of Civil Engineering, Purdue University. Email: sukkusur@purdue.edu

ABSTRACT

The spatial correlation between urban sprawl and the underlying road network has long been recognized in urban studies. The accessibility to road networks is often considered as an approximation for the measurement of human mobility, which is a key factor in determining potential urban sprawl in the future. Despite the close relationship between urban development and road networks, the spatial dependency of these two spatial layers has never been systematically evaluated. This paper conducts a comprehensive investigation on the spatial dependency between these two spatial layers using an urban expansion dataset between 2000 and 2010 of East Asian regions and the road network data from OpenStreetMap. Four Chinese cities, namely, Beijing, Shanghai, Chengdu and Shenzhen are selected to conduct the analysis. The spatial correlations between the urban sprawl and road networks are first quantitatively analyzed using Ripley's cross-K function. Highly significant spatial correlation have been observed in all the four tested cities. A Bayesian network model is also developed to verify the predictability of urban sprawl using the spatial and structural features extracted from the existing road networks as well as the spatial pattern of the past built-up areas. The results show an affirmative answer to the predictability of urban sprawl, by achieving an overall accuracy of 79% in classifying urban sprawl and undeveloped areas. Finally, the hidden dependencies among the urban sprawl and the extracted spatial features are interpreted and analyzed based on the Bayesian network structure learned from the data.

INTRODUCTION

The road network defines a basic template of the urban area that strongly constraints the urban development. It plays a prominent role in human mobility and activity analysis that directly impacts our understanding of urban sprawl patterns. The accessibility to road network is often considered as an important approximation for the measurement of human mobility, which is a key factor of the potential urban sprawl in a city (Obregón-Biosca et al. 2015). As a result, the co-location patterns of the existing urban settlement and the transportation infrastructures can be easily observed in many cities in the world. Such strong spatial correlation between urban built-up areas and road networks has naturally led to a hypothetical question: does strong spatial correlation exist between urban sprawl and road networks?

The evidences of the correlation between construction of new roads and future urban sprawl have long been recognized in literature (Harvey and Clark 1965; Bhatta 2010; Zischg et al. 2019). For example, the construction of highways leads to both congestion in the city and rapid outgrowth (Harvey and Clark 1965). A study by Yang (2002) on the urban sprawl of Atlanta during 1973-1999 also observed outward spread of high-density urban use along major transportation routes. Although the majority of urban sprawl studies mainly focus on analyzing macroscopic factors that impact the city growth, such as urban geometry, size relationship between cities, economic functions, social demographic and ethnic patterns, etc., many researchers have started incorporating road related features in modeling and forecasting urban sprawl (Yang and Lo 2003; Cheng and Masser 2003; Irwin and Bockstael 2007; Fregolent and Tonin 2016; Xu et al. 2014) due to the observation of existence of possible correlation between urban sprawl and road networks. For instance, Cheng and Masser (2003) developed a spatial logistic regression model to predict the urban sprawl pattern of the city of Wuhan in China. The model incorporated a set of variables that measure the distances of the given area to multiple types of road. Their study showed that the urban road infrastructure is one of the major determinants of urban growth.

Despite the growing consensus of the potential association of urban sprawl and the underlying road network, the spatial dependency patterns of these two spatial layers have never been systematically evaluated, and many important research questions still remain unanswered. For example, one simple but fundamental research question is that to what degree the urban sprawl and underlying road network are correlated. Quantifying the level of spatial correlation serves as a first step for us to fully understand the causal factors behind the coevolution of these two closely related spatial layers. Further, it helps to answer some interesting questions, such as: does a local cluster of urban sprawl always correspond to a local cluster of road networks? Are there any density threshold or other conditions exist on the structure and spatial distribution of the road network to enable the urban sprawl? If so, how can we design the road network to guide the desirable urban sprawl? This study aims to abridge these two spatial layers together by comprehensively investigating the spatial dependency between urban sprawl and road networks. Specifically, we investigate following sub-questions: (1) on what level the urban sprawl and road network are spatially correlated? (2) given the existence of such strong correlation, is urban sprawl predictable? And what are the relevant spatial features in determining the urban sprawl?

In this study, a large-scale urban expansion dataset between 2000 and 2010 of East Asia region from World Bank (Schneider et al. 2015; World Bank 2015) and the road network data extracted from OpenStreetMap (OpenStreetMap 2016) are used to conduct the analysis. Four different Chinese cities, namely, Beijing, Shanghai, Chengdu and Shenzhen are selected for the analysis. The spatial correlations between the urban sprawl and road networks are first quantitatively analyzed using cross-K function. The urban sprawl and the road network are converted into two spatial point processes with the original spatial relation information preserved to enable the utilization of cross-K function. A Bayesian network model is then developed to verify the predictability of urban sprawl using the spatial and structural features extracted from the existing road networks as well as the spatial patterns of the past built-up areas. Finally, the hidden relationships among the urban sprawl and the extracted spatial features are interpreted by analyzing the inferred structure of the Bayesian network.

The paper is organized as follows: the next section describes the data used in this study. Section 3 quantitatively evaluates the extent of spatial correlations between the urban sprawl and road networks of different cities. Section 4 develops an Bayesian network model to investigate the predictability of urban sprawl and further explores the contributing features associated with the urban sprawl. The final section concludes the paper.

DATA

The data used in this chapter were obtained from multiple sources. The urban sprawl data were obtained from a large-scale urban expansion dataset produced in a World Bank study (Schneider et al. 2015; World Bank 2015), which contains the urban expansion information across the East Asian region (stretching from Mongolia to the Pacific Islands) between 2000 and 2010. The data are in raster map format. The map is divided into 250×250m uniform cells and a specific label is assigned to each cell to indicate whether the cell is a built-up area before 2000, a new urban sprawl area between 2000 and 2010, or an undeveloped area by 2010. The road network data of the year 2012 from OpenStreetMap OpenStreetMap (2016) were extracted to provide an approximation of road network structure at the year 2010. Ideally, we want to use two sets of road network data around 2000 and 2010 to fully explore the spatial dependency between urban sprawl and road networks. Unfortunately, high-quality road network data in China around the year 2000 are not obtainable, hence we focus on analyzing the spatial dependency of urban sprawl and the road network structure after the urban sprawl.

Four representative cities from different geographical regions of China were selected to conduct the spatial dependency analysis, namely Beijing (north), Shanghai (east), Chengdu (southwest) and Shenzhen (south). All the four cities have experienced rapid urban sprawl during the 2000-2010 period. Both the urban sprawl and the road network data of the four cities were extracted. The urban sprawl pattern and the underlying road networks are illustrated in Fig.1a-1d (map data obtained from OpenStreetMap), and their summary statistics are presented in Table 1. The urban sprawl data in raster format were further processed and converted into a set of labeled points by placing a point at the center of each 250×250m cell. Every cell is thus represented by a specific point with a label that indicates the urban development condition of the area. The conversion from cell to points enables the urban sprawl to be modeled as a spatial point process and allows for more efficient computation in the analysis. For convenience, we refer to the points labeled as built-up areas as *built-up points*, the urban sprawl area as *urban sprawl points* and the rest of the points as *undeveloped points*.

SPATIAL CORRELATION BETWEEN URBAN SPRAWL AND ROAD NETWORK

The co-location patterns of urban sprawl and road networks are evident in many cities, since roads are built to make the human settlements accessible to other parts of the city. However, such correlation patterns have never been systematically investigated, and many research questions still remain. For example, to what degree these two spatial layers are correlated? Does a local cluster of urban sprawl always correspond to a local cluster of road networks? To adequately address these questions, the spatial correlation between the urban sprawl and road network needs to be quantitatively evaluated. In this section, we develop a quantitative evaluation method to examine the degree of spatial correlation between urban sprawl and road networks. The approach is based on a new distance measure (point-to-road distance) for evaluating the spatial proximity between a point process and a spatial network, as well as the cross-K function.

Model development

There is a large body of research and statistical techniques on measuring the spatial correlation between two spatial point processes, including the Ripley's Cross-K function with Monte Carlo simulation (Cressie 1993), mean nearest-neighbor distance (Dixon 2002), and spatial regression models (Chou 1997). However, there is no existing method on evaluating the spatial correlation between a point process and a spatial network, as the spatial relationship between a point and a line segment is far more complex than a pair of points. To utilize the well-established theoretical results on the spatial correlations of point processes as well as generalize the analysis to point-network analysis, we construct a new point process from the road network while preserving sufficient spatial relationship between the given point process and the original road network. The key step of this construction is to introduce a new set of points (referred as *access point*) and a new distance measure (referred as *point-to-road distance*) to evaluate the spatial proximity between a point and a road segment. Given a target point (e.g. an urban sprawl point), the corresponding access point on a road segments is defined as its projection (point on the road segment that has the minimum distance to the target point) on the road segment. Since the points on a curve does not necessarily form a convex set, the projection operation may not lead to the unique solution. Under such cases, we only pick one point in the solution set as the access point. The point-to-road distance between the target point and a road segment can thus be defined as the great-circle distance (the shortest distance between two points on the surface of a sphere) with Earth radius between the target point and the corresponding access point of the road segment. Fig.2 presents a conceptual illustration of the aforementioned access point and the point-to-road distance. There are three important features of this construction:

1. Every target point has only one access point for each road segment.
2. The point-to-road distance captures the spatial proximity of the target point and a road segment. As a large point-to-road distance indicates all points on the road segment are far away from the target point.
3. All roads that have a point-to-road distance smaller than h to the target point will have non-empty intersection with the area formed by the neighborhood of the target point with radius h. Hence the set of access point in the neighborhood of the target point corresponds to the same set of road segments that intersect with the neighborhood area.

Given the above three important features of the access points and the point-to-road distance, the spatial relationship between the point process and the spatial network can be adequately captured. The spatial correlation of the two spatial layers can thus be approximated by the spatial correlation between the point process and a set of constructed access point processes for each target point in the original point process.

The most widely used spatial statistics for evaluating the spatial correlation between two point processes is the cross-K function, a generalization of Ripley's K-function (Cressie 1993; Huang et al. 2004; Dixon et al. 2002). The cross-K function for binary spatial features is defined as follows:

$$K_{ij}(h) = \lambda_j^{-1} E[\text{number of type } j \text{ instances within distance of a randomly chosen type } i \text{ instance}]$$

where λ_j is the density (number per unit area) of type j instances and h is the distance. Without edge effect (Cressie 1993; Dixon et al. 2002), the cross-K function can be estimated by

$$\hat{K}_{ij}(h) = \frac{1}{\lambda_i \lambda_j A} \sum_k \sum_l I_h(d(i_k, j_l)) \tag{1}$$

where A is the total area of the study region, $d(i_k, j_l)$ is the distance between the kth instance of type i and the lth instance of type j; $I_h(d(i_k, j_l))$ is the indicator function which takes value 1 if $d(i_k, j_l)$ is smaller than h, 0 otherwise. The edge effect arises because points outside the study region are not counted in the numerator, even if they are within distance h of a point in the study region (Dixon et al. 2002). However, fully removal of the edge effect needs to perform the computationally expensive edge-correction, thus is often omitted (Cressie 1993). In our case, since the spatial correlation is measured between the urban sprawl points and their corresponding access point sets, above estimator needs to be modified to make it suitable for our analysis. Denote $P = \{p_i, i = 1, 2, \cdots, n\}$ as the set of urban sprawl points, $Q = \{q_r^i, r = 1, 2, \cdots, m\}$ as the set of access points of urban sprawl point p_i, in which q_r^i is the access point of p_i on road r, and $d_{pr}(p_i, r) = d(p_i, q_r^i)$ is the point-to-road distance between point p_i and road r. The equivalent estimator of cross-K function without edge-correction can be modified as follows

$$\hat{K}_{ij}(h) = \frac{A}{nm} \sum_{i=1}^{n} \sum_{r=1}^{m} I_h(d_{pr}(p_i, r)) \tag{2}$$

The cross-K function characterizes pairwise spatial relationship between two point processes. To test the extent of spatial correlation between two point processes, a typical treatment is to compare the results against the curve of $K(h) = \pi h^2$, which indicates complete spatial randomness when edge effect is not present. However, in this study, the edge effect is present. This can be easily observed from the urban sprawl pattern of Shanghai (Fig.1b) and Shenzhen (Fig.1d), as a considerable amount of urban sprawl areas are located near the boundary of the two cities. Thus comparing against the curve of $K(h) = \pi h^2$ is not appropriate and may lead to biased results. In this study, we adopt an alternative approach by comparing the result against a baseline cross-K function produced by the urban sprawl points and a completely independent point process with points randomly generated inside the study region. The random points generation is performed using Monte Carlo simulation with the number of random points equal to the number of road segments. We run the simulation 20 times for each city and report the 0.05-0.95 quantile of the computed baseline cross-K functions.

Experimental results

Fig.3a-3d show the results of the cross-K function for the four cities investigated in this study. In all of the four cities, the curves of cross-K functions are much higher than the 0.05-0.95 quantile of the baseline cross-K functions. It suggests that in all distance scales, the average number of road segments within distance h to an urban sprawl area is much higher compared with an independent and randomly distributed point process with the same density. This clearly indicates that strong attraction behavior exists, that the urban sprawl areas and the road segments tend to be co-located. The cross-K function results for Beijing, Shanghai and Shenzhen exhibit similar patterns, where their cross-K function values are about twice as much as the baseline cross-K function values, which means in average, there are about twice the number of road segments located within certain distance to an urban sprawl point compared with the number of points generated by an independent randomly distributed spatial process. The variance of the baseline cross-K function of Shenzhen is a little larger (wider 0.05-0.95 quantile range) compared with the results of Beijing and Shanghai, which might caused by more significant edge effect. Different from the previous three cities, Chengdu exhibits much stronger spatial correlation between urban sprawl and road networks, as its cross-K function curve is significantly higher than the baseline cross-K function curve. This might caused by the mountainous terrain surrounding Chengdu, that most of the urban development areas and roads are concentrated in a series of clusters. The higher level of local clustering contributes to a stronger spatial correlation between the two spatial layers. Despite the differences, the results in all the four cities confirmed the existence of strong spatial correlation between the urban sprawl and the underlying road networks.

PREDICTABILITY OF URBAN SPRAWL GIVEN UNDERLYING ROAD NETWORK

The previous section provides an affirmative answer to the existence of strong correlation between urban sprawl and the underlying road network. Given the existence of strong spatial correlation, a natural and more interesting research question to be answered is that: is urban sprawl predictable using the information contained in the road network structures? And what are the relevant spatial and structural features in the road network that provide discriminative information on deciding whether an area is part of the future urban sprawl or remains undeveloped during a specific observation period. Answering these questions will not only contribute to a better understanding of the mechanism behind the co-evolution of urban sprawl and road network growth, but also have important practical implications. For example, understanding of the structural impacts of road networks can provide insights on guiding road network construction that lead to desirable and healthy urban sprawl in the future.

In this section, we focus on addressing aforementioned questions by verifying the predictability of the urban sprawl using spatial and structural features extracted from the road network as well as the spatial pattern of the past built-up areas. The problem is cast into a binary classification and prediction problem on the area label (urban sprawl or undeveloped). A Bayesian network model is developed to learn the dependencies and causal relationships between extracted spatial features and the area labels, and further predicts the target area label.

Spatial feature extraction

To begin our analysis, a set of spatial features were first extracted from the urban sprawl and the road network data. A list of extracted features as well as their descriptions are presented in Table 2. Our target variable is *sprawlLabel* which encodes the actual state of the area, e.g. whether it is an urban sprawl or undeveloped area. Two global statistics are computed, namely *minUrbDist* and *minRoadDist* which are the distances of the area to the nearest built-up area and the road segment (using point-to-road distance) respectively. Local spatial features within 1km radius of a target area were also extracted, including the number of built-up areas (*NneighUrb*) and road segments (*NneighRoad*) fall within the 1km radius; the number of road intersections with 3 or more approaches (*Nintersection*); and the lengths of each road type based on corresponding type information provided in OpenStreetMap (*primaryLen, secondaryLen, localLen, otherLen*). More detailed information on the extracted features can be found in Table 2. When computing the road lengths, we only calculated the part of length of the road segment within the 1km radius circle, rather than the total length of the road segments. The local road density can thus be obtained by summing the extracted road length of different types and divided by $\pi \times 1^2 km^2$.

Since the feature extraction is computationally extensive, for each city, we randomly selected 500 urban sprawl areas and 500 undeveloped areas as instances and extracted all the spatial features for every instance. Hence, for each of the four cities, we obtain a dataset of 1000 instances. After that, the dataset for each city were randomly partitioned into 2 sub-datasets, the first contains 800 instances which served as the training set, and the second contains 200 instances which served as the testing set. In the actual implementation of the urban sprawl prediction model, the feature *otherLen* was removed from the model training and testing phase. As it is correlated with *primaryLen, secondaryLen, localLen* and *roadDensity* ($primaryLen + secondaryLen + localLen + otherLen \propto roadDensity$) and does not provide additional information.

Urban sprawl prediction

Since all instance labels are known, we adopted the supervised learning approach in machine learning to predict the area labels. Four widely used supervised learning classifiers were tested and examined in this study, namely, Naïve Bayes, support vector machine (SVM), random forest and Bayesian network. The Naïve Bayes is a simple probabilistic classifier based on Bayes' Theorem, which assumes conditional independence among features. SVM is another popular method, which classifies data by maximizing the distance between the decision boundaries defined by a set of control points (support vectors) (Bishop 2006). On the other hand, the random forest (Breiman 2001) approach is an ensemble method, which uses a combination of randomly generated decision tree classifiers to increase accuracy. The Bayesian network (Bishop 2006) is a probabilistic graphical model which represents a set of random variables $U = \{x_1, x_2, \cdots, x_n\}, n \geq 1$ and their conditional dependencies via a directed acyclic graph (DAG). A Bayesian network represents a probability distribution as follows

$$P(U) = \prod_{u \in U} p(u|pa(u)) \tag{3}$$

where $p(u|pa(u))$ is the conditional probability table between random variable u and its parents $pa(u)$. More detailed background information about the theory, learning and inference procedures of Bayesian network can be found in (Bishop 2006; Koller and Friedman 2009). We used the machine learning software Weka (Hall et al. 2009) to implement all of the previous four supervised learning classifiers. The training dataset for all of the four cities were combined (4000 instances) to train each classifier, and the overall classification accuracy on the combined test datasets (800 instances) were used as the final criteria for model selection. The Bayesian network was found to achieve the highest overall classification accuracy of 79%, thus was selected as the final model to perform the urban sprawl prediction task. Another advantage of using the Bayesian network for our analysis is that Bayesian network has the ability to reflect causal relationships between between variables. As is shown in Pearl (Pearl 2009), if a set of variables have causal relations, and the Bayesian network is built such that arcs fully represent the causal paths between variables, then the resulting Bayesian network will encode dependencies and probabilistic relations between variables. This property is particularly helpful for us to identifying relevant spatial features that impact the urban sprawl.

To implement the Bayesian network, we first discretized each continuous feature into a set of discrete states. The K2 algorithm developed by Cooper et al. (Cooper and Herskovits 1992) was then applied to learn the structure of the Bayesian network as well as the parameters in the conditional probability tables $p(u|pa(u)), \forall u \in U$. In the prediction phase, using the already learned conditional probability tables, the area label for area i is predicted as follows:

$$sprawlLable_i^* = argmax_{sprawlLable} P(sprawlLable|\mathbf{x}^i) \propto \prod_{u^i \in U} p(u^i|pa(u^i)) \tag{4}$$

where $\mathbf{x}^i$ is the set of all observed features except *sprawlLable* for area i.

Experimental result

The structure of the Bayesian network learned from the training data is presented in Fig. 4. It can be observed that all the extracted spatial features are closely associated with the area label, as there is a direct arc connecting *sprawlLable* to every other feature. This indicates that the emergence of an urban sprawl area has the potential to impact all the local features of the road network. It confirms our intuition that urban sprawl and road network are co-evolved with each other, and the development of urban settlement will lead to the construction of new roads in order to establish connectivity to other parts of the city. Among other spatial features, the conditional dependency between the distance to closest road segments (*minRoadDist*) and the road density (*roadDensity*) is obvious, since a location that is far away from any other road segments is not possible to have high local road density. The road lengths of different types (*primaryLen, secondaryLen, localLen*) are conditionally dependent on road density (*roadDensity*) and the number of neighboring road segments (*NneighRoad*), which is also intuitive. The distance to the closest road segment (*minRoadDist*) and the number of neighboring road segments (*NneighRoad*) are conditionally dependent on the distance to the closest built-up area, which also makes sense. As locations that are far away from any existing built-up areas are less likely to have dense road network as well as the potential of urban sprawl. From above analysis, it can be observed that the structure of Bayesian network provides rich information about the hidden relationship among the extracted spatial features, and allows us to better understand the role of each spatial feature in the emergence of urban sprawl.

The developed Bayesian network model was validated using the test dataset of the four studied cities. The testing results of the combined test set as well as the test set for each city are presented in Table 3. The confusion matrix for the experiment using the combined test set is presented in Table 4. In addition to the accuracy measure, F-measure (also refer as F1) is used to evaluate the classification quality on the test sets, which is a commonly used accuracy measure in data mining. F-measure is computed as the harmonic mean of precision (percentage of testing data that are classified as positive are actually positive) and recall (percentage of positive testing data that are classified as positive), which is calculated as $\textit{F-measure} = 2(precision \times recall)/(precision + recall)$. A higher F-measure suggests a better classification result.

It can be observed that even only use the spatial information provided in road networks as well as the past built-up areas, the Bayesian network can achieve about 80% overall accuracy in classifying the area labels. The accuracy of both Beijing and Shanghai are around 80%. For the best case - the result of Chengdu, the accuracy even achieved 85%. As we have already observed in Section 3, Chengdu exhibits highest level of spatial correlation between urban sprawl and the road network. This confirmed our hypothesis that high level of spatial correlation indeed contributes to the predictability of urban sprawl. The test accuracy of the experiment for Shenzhen is the lowest, however, still reaches 73%. The relatively lower accuracy for Shenzhen might again associated with the edge effect discussed in Section 3, which also impacts the spatial feature extraction. As the training and testing instances are randomly selected, if they are located close to the boundary of the study region, the local spatial features will be incomplete and lead to inaccurate information. For the results of the F-measure, not surprisingly, Chengdu has the highest F-measure for both area label classes among the four cities. The F-measure for urban sprawl areas are consistently higher than the F-measure for areas remain undeveloped, which suggests lower misclassification error in predicting urban sprawl areas, this partly indicates the urban sprawl is relatively more predictable than the areas that will remain undeveloped during the observation period. All the results show that the urban sprawl is highly predictable given the strong spatial correlation between urban sprawl and the road networks. Furthermore, the dependency relationship among the extracted spatial features revealed by the structure of Bayesian network can also be used as an input to guide the road network construction that leads to a more desirable and healthy urban sprawl in the future.

CONCLUSION

This paper systematically investigates the spatial dependency between urban sprawl and the underlying road networks. An urban sprawl dataset between 2000 and 2010 for East Asia region and the road networks from OpenStreetMap are used in this study. Four Chinese cities, namely, Beijing, Shanghai, Chengdu and Shenzhen are selected to conduct the analysis. The spatial correlation between the urban sprawl and road network is first quantitatively evaluated using cross-K function. Highly significant spatial correlation has been observed in all of the four tested cities. A Bayesian network model is then developed to verify the predictability of urban sprawl given the existence of strong spatial correlation of the two spatial layers. The results provide an affirmative answer to the predictability of urban sprawl, that about 79% of overall prediction accuracy is achieved by using a set of spatial and structural features extracted from the road networks as well as the spatial patterns of the past built-up areas.

There are some limitations of this study. First, since the historical road networks around year 2000 is not obtainable, some of the research questions related to the co-evolution of the urban sprawl and road network can not be fully explored. Also, as this study focuses on exploring the spatial dependencies between the urban sprawl and the road network structure, thus the prediction of urban sprawl solely rely on the spatial correlation of these two spatial layers. It is expected that even higher prediction accuracy can be achieved when incorporating additional information, e.g. the terrain feature of the city, demographics characteristics such as population growth, GDP growth during the observation period and the prior knowledge from the local planning policies, etc. Future research can be done to develop more comprehensive decision support tools for city planners for more accurate urban sprawl prediction.

REFERENCES

Bhatta, B. (2010). *Analysis of urban growth and sprawl from remote sensing data*. Springer Science & Business Media.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Breiman, L. (2001). “Random forests.” *Machine learning*, 45(1), 5–32.

Cheng, J. and Masser, I. (2003). “Urban growth pattern modeling : a case study of Wuhan city , PR China.” *Landscape and urban planning*, 62(4), 199–217.

Chou, Y.-H. (1997). “Exploring spatial analysis in geographic information systems.” *Exploring Spatial Analysis in Geographic Information Systems*, OnWord Press.

Cooper, G. F. and Herskovits, E. (1992). “A bayesian method for the induction of probabilistic networks from data.” *Machine learning*, 9(4), 309–347.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, Vol. 14 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Hoboken, NJ, USA (September).

Dixon, P. M. (2002). “Nearest neighbor methods.” *Encyclopedia of environmetrics*.

Dixon, P. M., El-shaarawi, A. H., and Piegorsch, W. W. (2002). “Ripley’ s K function.” *Encyclopedia of Environmetrics*, 3, 1796–1803.

Fregolent, L. and Tonin, S. (2016). “Local public spending and urban sprawl: Analysis of this relationship in the Veneto region of Italy.” *Journal of Urban Planning and Development*, 142.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). “The weka data mining software: an update.” *ACM SIGKDD explorations newsletter*, 11(1), 10–18.

Harvey, R. O. and Clark, W. A. (1965). “The nature and economics of urban sprawl.” *Land Economics*, 1–9.

Huang, Y., Shekhar, S., and Xiong, H. (2004). “Discovering colocation patterns from spatial data sets: a general approach.” *IEEE Transactions on Knowledge and Data Engineering*, 16(12), 1472–1485.

Irwin, E. G. and Bockstael, N. E. (2007). “The evolution of urban sprawl: evidence of spatial heterogeneity and increasing land fragmentation.” *Proceedings of the National Academy of Sciences*, 104(52), 20672–20677.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Obregón-Biosca, S. A., Romero-Navarrete, J. A., Mendoza-Sanchez, J. F., and Betanzo-Quezada, E. (2015). “Impact of mobility induced by urban sprawl: Case study of the Querétaro metropolitan area." *Journal of Urban Planning and Development*, 142(2), 05015005.

OpenStreetMap (2016), <https://www.openstreetmap.org/>.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Schneider, A., Mertes, C. M., Tatem, A. J., Tan, B., Sulla-Menashe, D., Graves, S. J., Patel, N. N., Horton, J. A., Gaughan, A. E., Rollo, J. T., Schelly, I. H., Stevens, F. R., and Dastur, A. (2015). "A new urban landscape in east–southeast asia, 2000-2010." *Environmental Research Letters*, 10(3), 034002.

World Bank (2015). "World bank urban spatial dataset: Puma, <http://puma.worldbank.org/downloads>.

Xu, C., Fang, S., Long, N., Teng, S., Zhang, M., and Liu, M. (2014). "Spatial patterns of distinct urban growth forms in relation to roads and pregrowth urban areas: Case of the nanjing metropolitan region in china." *Journal of Urban Planning and Development*, 141(1), 04014015.

Yang, X. (2002). "Satellite monitoring of urban spatial growth in the Atlanta metropolitan area." *Photogrammetric Engineering and Remote Sensing*, 68(7), 725–734.

Yang, X. and Lo, C. P. (2003). "Modelling urban growth and landscape changes in the Atlanta metropolitan area." *International Journal of Geographical Information Science*, 17(5), 463–488.

Zischg, J., Klinkhamer, C., Zhan, X., Rao, P. S. C., and Sitzenfrei, R. (2019). "A century of topological coevolution of complex infrastructure networks in an alpine city." *Complexity*, 2019, 2096749.

List of Tables

1 Summary of the urban sprawl in the four cities 18

2 Description of the extracted features 19

3 Testing results for the Bayesian network 20

4 Confusion matrix for the experiment using combined testing set 21

TABLE 1. Summary of the urban sprawl in the four cities

Summary statistics	Beijing	Shanghai	Chengdu	Shenzhen
Total area (km^2)	9,910.31	5,264.28	11,487.77	1,858.72
Total number of built-up points	28,546	17,549	11,113	9,597
Total number of urban sprawl points	13,992	16,428	8,484	4,099
Total number of undeveloped points	116,044	50,259	16,4232	16,048
Total number of road segments	46,711	11,151	47,442	18,739
Population at 2000	1633.0	1858.0	1257.9	701.2
Population at 2010	1961.2	2301.9	1404.8	1035.79
Population growth	20.10%	23.89%	11.68%	47.71%

TABLE 2. Description of the extracted features

Feature name	Description
sprawlLabel	Binary variable indicating the class of the $250 \times 250m$ area: urban sprawl (1) and undeveloped area (0)
minUrbDist	Great circle distance (km) to the nearest built-up area
minRoadDist	Point-to-road distance (km) to the nearest road segment
NneighUrb	Number of existing built-up areas within 1km radius
NneighRoad	Number of road segments within 1km radius
roadDensity	Road density within 1 km radius (km^{-1})
Nintersection	Number of intersections with at least 3 approaches within 1km radius
primaryLen	Total distance (km) of primary road (e.g. primary link, motorway, raceway)
secondaryLen	Total distance (km) of secondary road (e.g. secondary link)
localLen	Total distance (km) of local road (e.g. residential, services, pedestrian, footway)
otherLen	Total distance (km) of other road types in OpenStreetMap

TABLE 3. Testing results for the Bayesian network

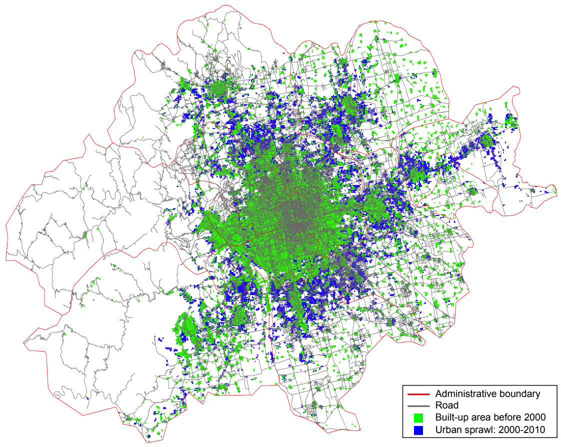
Testing set	Beijing	Shanghai	Chengdu	Shenzhen	Combined
Number of test instances	200	200	200	200	800
F-measure: urban sprawl area	0.818	0.814	0.838	0.733	0.799
F-measure: undeveloped area	0.778	0.732	0.865	0.727	0.780
Overall accuracy	80%	78%	85%	73%	79%

TABLE 4. Confusion matrix for the experiment using combined testing set

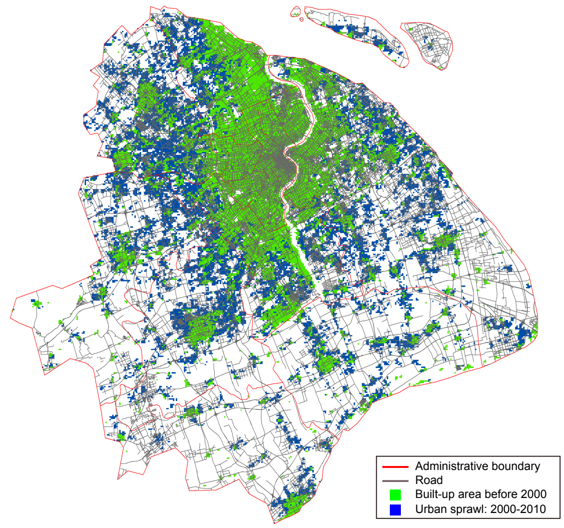
Actual/Predicted	Urban sprawl	Undeveloped area	sum
Urban sprawl	334	80	414
Undeveloped area	88	298	386
Overall accuracy	79%		

List of Figures

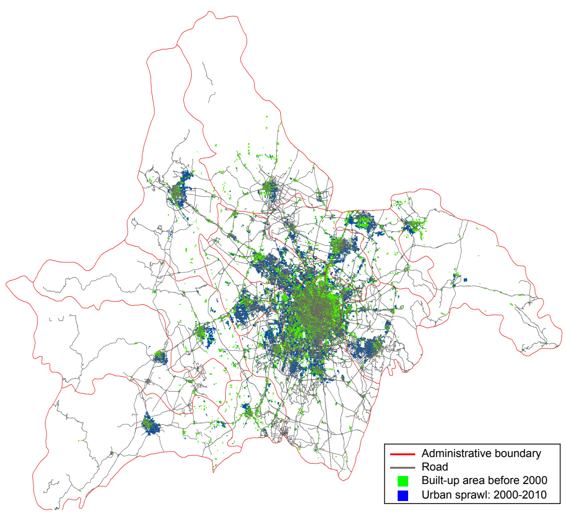
1 Illustration of urban sprawl of four Chinese cities from 2000 to 2010 and their corresponding road networks . . . 23
1(a) Beijing . . . 23
1(b) Shanghai . . . 23
1(c) Chengdu . . . 23
1(d) Shenzhen . . . 23
2 Conceptual illustration of the access point and the point-to-road distance . . . 24
3 Cross-K function for four cities in China. Solid lines are the plots of cross-K function $K(h)$. Dash lines are 0.05 and 0.95 quantiles of the baseline cross-K function of the urban sprawl points and a randomly generated point process estimated from 20 Monte Carlo simulations. . . . 25
3(a) Beijing . . . 25
3(b) Shanghai . . . 25
3(c) Chengdu . . . 25
3(d) Shenzhen . . . 25
4 The structure of the Bayesian network learned from data . . . 26



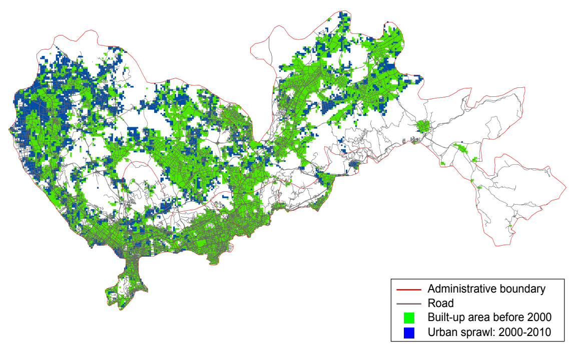
(a) Beijing



(b) Shanghai



(c) Chengdu



(d) Shenzhen

Fig. 1. Illustration of urban sprawl of four Chinese cities from 2000 to 2010 and their corresponding road networks

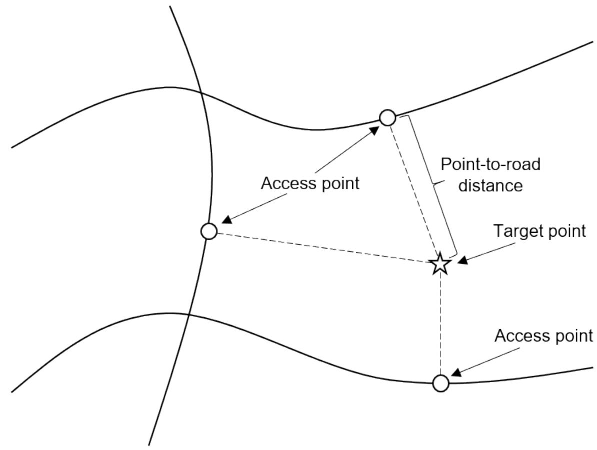
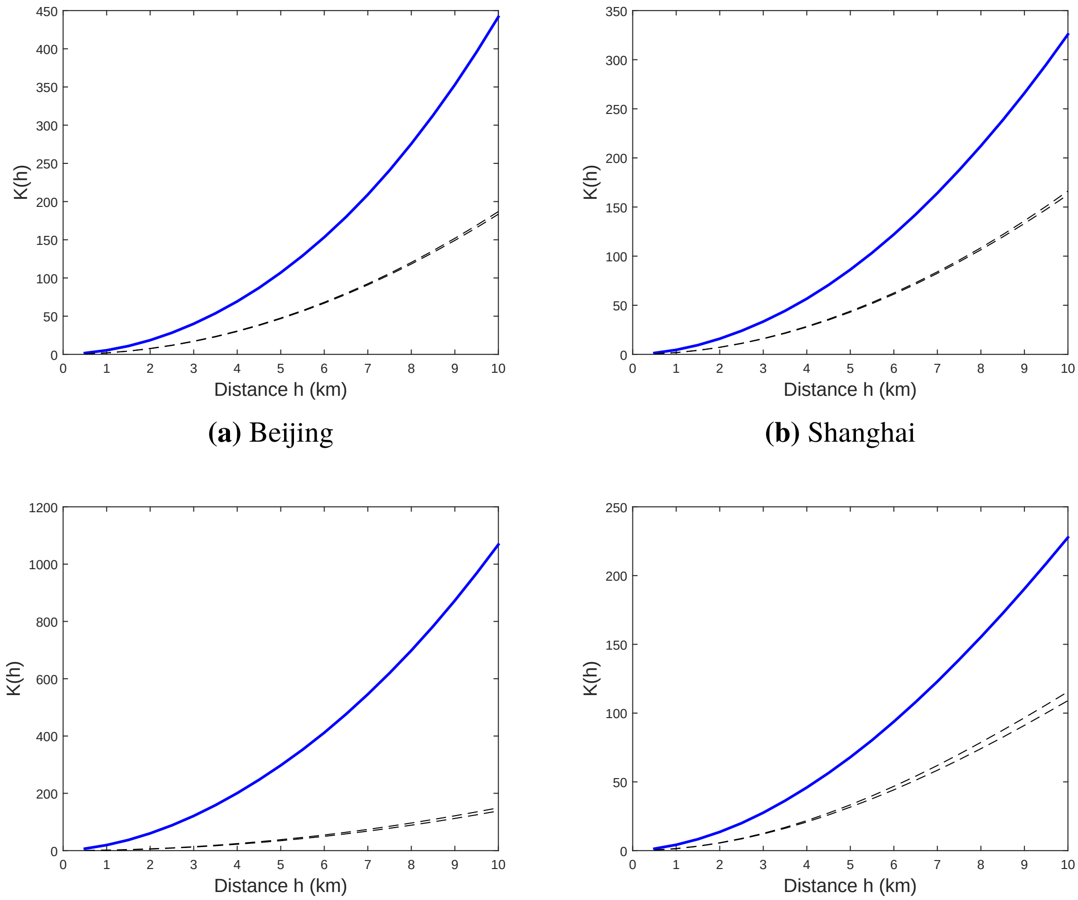


Fig. 2. Conceptual illustration of the access point and the point-to-road distance



(a) Beijing **(b)** Shanghai **(c)** Chengdu **(d)** Shenzhen

Fig. 3. Cross-K function for four cities in China. Solid lines are the plots of cross-K function $K(h)$. Dash lines are 0.05 and 0.95 quantiles of the baseline cross-K function of the urban sprawl points and a randomly generated point process estimated from 20 Monte Carlo simulations.

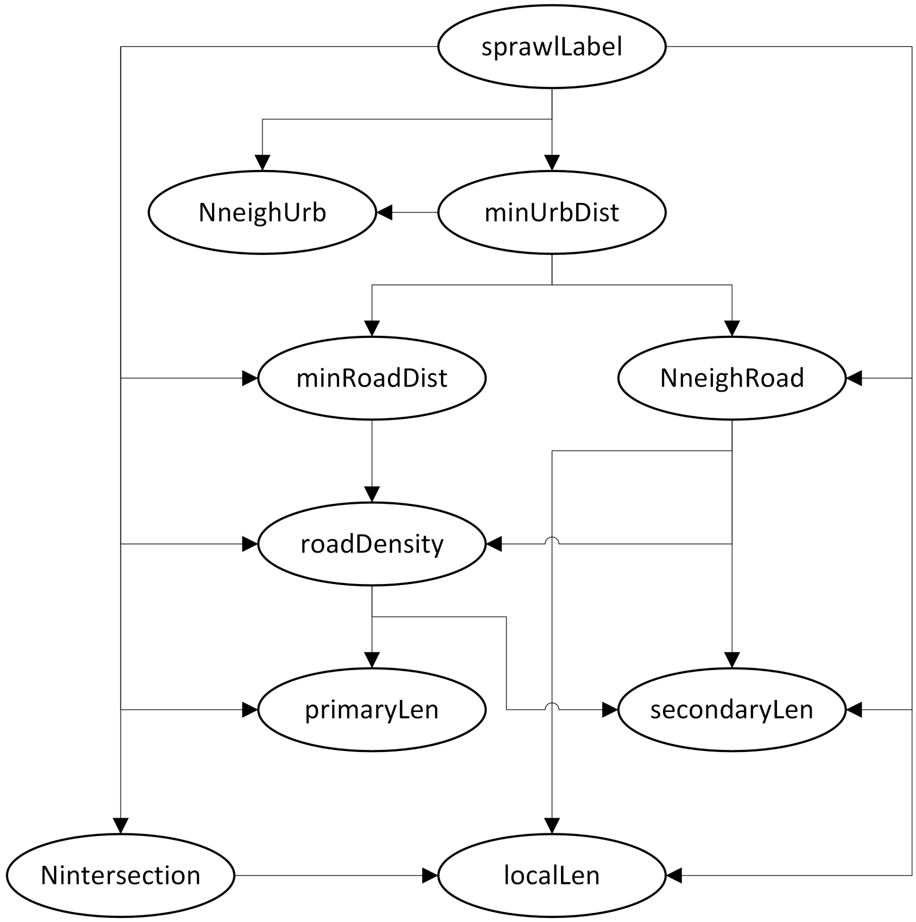


Fig. 4. The structure of the Bayesian network learned from data