Geo-propagation from Incomplete Spatial Distribution Data: A Case Study of House Price Estimation

Di Zhu[*1,2], Tao Cheng[†1], Yu Liu[‡2]

[1]University College London (UCL), Department of Civil, Environmental & Geomatic Engineering, London, United Kingdom
[2] Peking University, Institute of Remote Sensing and Geographic Information Systems, Beijing, China

Jan 18, 2019

Summary

Despite the emergence of big data that contains plentiful spatiotemporal information, researchers are still facing the challenge that the acquired data after time slicing are insufficient to characterize the real heterogeneous spatial pattern. The problem becomes especially important in areas such as epidemiology and economy where the fine-resolution data are under intense demand. This paper introduces a novel approach named *geo-propagation* to propagate sampled data to unsampled locations incorporating semi-supervised learning under the graph context. A case study of estimating property price is shown to demonstrate the effectiveness of our proposed method.

KEYWORDS: Spatial estimation, semi-supervised learning, graph, propagation, house price

1. Backgrounds

When representing and understanding a geographical phenomenon, such as the spatial distribution of precipitation, we are often forced to collect a limited number of samples instead of observing at every possible location (Goodchild et al. 1993). Spatial estimation is an operation that generally aims at predicting the value $\mathrm{z}(x^*)$ at an unobserved location x^* given a sample of data $\mathrm{z}(x_i)$, $i = 1, 2, \dots, m$ (Atkinson and Lloyd 2009).

Current methods of estimating from incomplete spatial distribution data are based on the spatial dependency assumption indicated in Tobler's First Law of Geography (Tobler 1970), or the local heterogenous conditions indicated by Goodchild (2004b). All these methods require an increasing number of samples and a balanced sampling strategy to improve the accuracy. The inference is based on either global fitted models or local regression models. In other words, previous spatial estimation methods mainly adopted the supervised learning system to train predictive models.

However, in areas such as house price and public health, there is hardly a good spatial coverage of the data after time slicing. The limited historical survey data from priori locations is often insufficient for training a good model under the supervised learning context. In this study, we propose *geo-propagation*, an idea that formalizes locations into connected graphs and then uses semi-supervised propagation algorithms to achieve spatial estimation. The major contribution is to consider both local variations and global dependences from a network perspective and to simulate the propagation of sampled data in space.

2. Methodology

We will first present common ways to create graphs that connect locations. Then we will clarify the

* patrick.zhu@pku.edu.cn (correspondence)
† tao.cheng@ucl.ac.uk
‡ liuyu@urban.pku.edu.cn

mathematical implementation of geo-propagation.

2.1. Common Ways to Create Graphs that Connect Spatial Locations

In perspective of the First Law of Geography (Tobler 1970), we can simply use a distance decay function to model the spatial displacement (Zhu 2018). For example, consider a variant of the self-tuning Gaussian diffusion kernel:

$$W_{ij} = exp^{-\frac{d(i,j)}{\sigma_i \sigma_j}} \tag{1}$$

where $d(i,j)$ is the Euclidean distance between node i and j and σ_i is computed as the distance $d(i,i_k)$ corresponding to the *k-th* nearest neighbor i_k of node i. Also, we can model the connections in a gravity model where c, α, β and γ are preset coefficients:

$$W_{ij} = c\frac{P_i^{\alpha} P_j^{\beta}}{d(i,j)^{\gamma}} \tag{2}$$

On the other hand, considering the Third Law of Geography that "The more similar geographic configurations of two locations, the more similar the values of the target variable at these two locations" (Zhu 2018), we can project locations from Euclidean space into a higher dimension space formalized by covariates. An example of using three covariates to quantify the similarity between locations' configurations is shown in Figure 1.

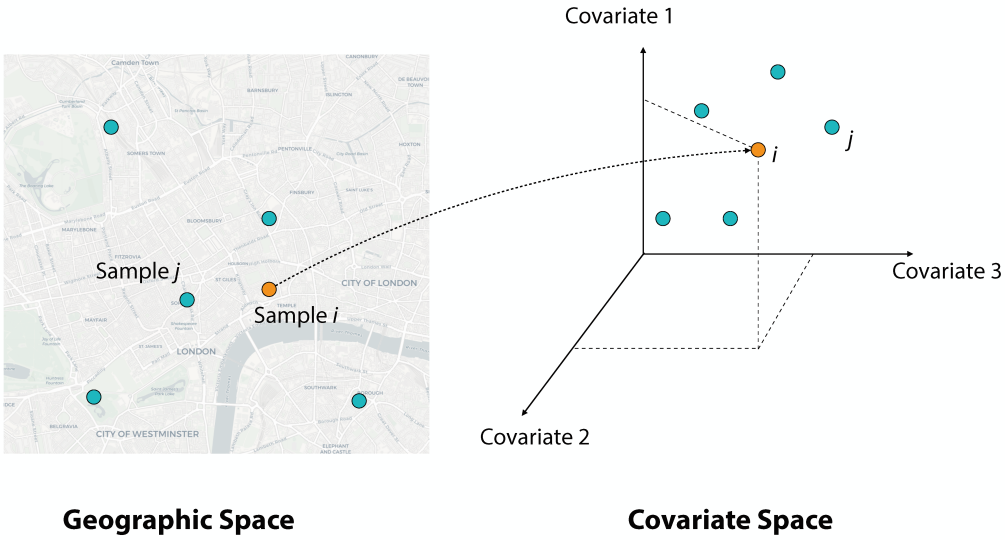

Figure 1 Calculating the similarity of geographical configurations between locations.

Intuitively we want locations that are closer in covariate space to have similar values, a direct definition of similarity network can be:

$$W_{ij} = exp\left(-\frac{||L_i - L_j||^2}{\alpha^2}\right) \tag{3}$$

where α is a bandwidth hyperparameter, L_i is the coordinates of location i.

In practice, a weighted combination of Equations (1), (2), (3) and their variants may be used together by domain experts to obtain a good graph structure.

2.2. Convergence of Geo-propagation

With the definition of graphs, we propagate the sampled values through the edges based on the assumption that a high edge weights allow values to transmit more easily. This is the basic idea of geo-propagation. An $n \times n$ transition matrix P can be defined as follow:

$$P_{ij} = P(i \to j) = \frac{W_{ij}}{\sum_{k=1}^{n} W_{ik}} \tag{4}$$

For spatial estimation problems, we have s sampled locations as a ground truth $s \times C$ matrix Y_S, where C is the number of target variables, namely one in a simple case. The u unsampled locations have a soft label matrix f_U to calculate and the s sampled locations also have a soft label matrix f_S accordingly ($s + u = n$). The objective of geo-propagation is to compute f_U.

Let $f = \begin{pmatrix} f_S \\ f_U \end{pmatrix}$ and rewrite P into sampled and unsampled sub-matrices $P = \begin{bmatrix} P_{SS} & P_{SU} \\ P_{US} & P_{UU} \end{bmatrix}$, we propagate the values in the graph by keep computing a new $f^{(i+1)} \leftarrow Pf^{(i)}$ and resetting $f_S = Y_S$ until f_U convergence. Since that $f_U \leftarrow P_{US}Y_S + P_{UU}f_U$, after k times of iteration, we have

$$f_U^{(k)} = (P_{UU})^k f_U^{(0)} + \left[\sum_{i=1}^{k} (P_{UU})^{i-1}\right] P_{US}Y_S \tag{5}$$

According to Equation (4), $\lim_{k\to\infty} (P_{UU})^k f_U^{(0)} = 0$. Thus, the initial value $f_U^{(0)}$ is inconsequential and the analytical convergence solution of Equation (5) can be obtained as

$$f_U = \lim_{k\to\infty} \left[\sum_{i=1}^{k} (P_{UU})^{i-1}\right] P_{US}Y_S = (I - P_{UU})^{-1} P_{US}Y_S \tag{6}$$

as long as every connected component in the graph has at least one sampled location (so that $I - P_{UU}$ is invertible). Similar convergence has been demonstrated by Zhu (2005) in a label context instead of the numerical data propagation.

Since the number of unsampled spatial location can be very huge in practice, using iteration to compute the optimal result often outperforms Equation (6) as the computing the inverse of $I - P_{UU}$ can be time-consuming. In addition, there are actually many other ways to approximate the convergence, such as graph convolutional neural networks, we won't discuss them due to the page limit.

3. Case Study

Variations of house prices in space and time have significant impact on household welfare, financial stability and urban planning. An accurate house price forecasts is therefore important for government and real estate industries. However, previous research is reluctant in practice either due to the sparse price information after temporal slicing nor the overlook of geographic configuration in models.

3.1. Data descriptions

We utilized a dataset of traded housing price in *Beijing* from 2011 to 2016 fetched from *https://bj.lianjia.com/chengjiao/* to test the feasibility of geo-propagation. The number of sold properties gradually increases from 2011 to 2016 (Figure 2a), showing an active trading market in Beijing with a high demand of accurate price estimations. Figure 2b indicates both the number and price of traded properties have increased a lot during the five years. Also, we can continuously see newly built residential communities each year whose property prices need to be evaluated (Figure 2c).

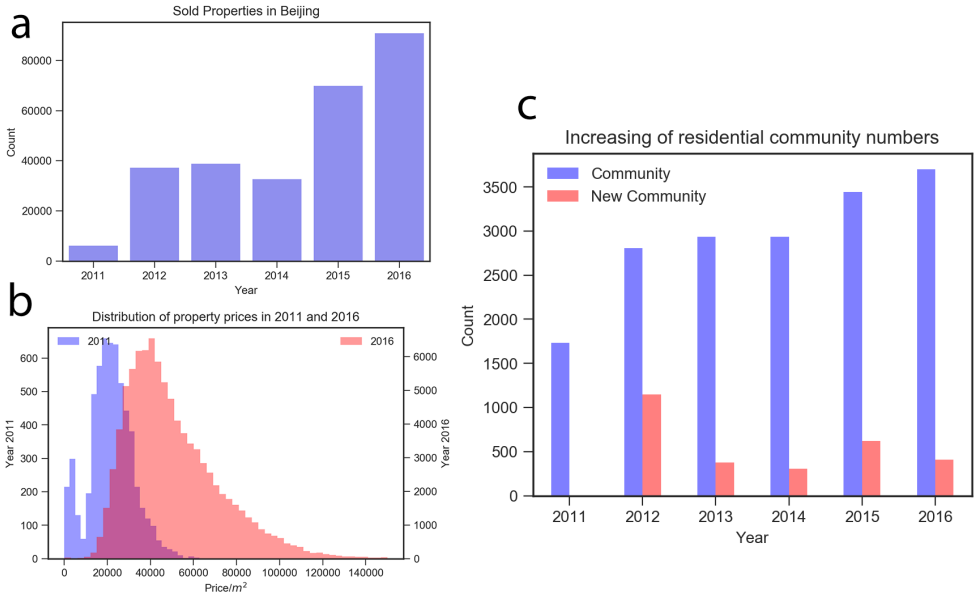

Figure 2 Statistics of the house price market in *Beijing* from 2011 to 2016 (a) number of sold properties. (b) histogram of prices in 2011 and 2016. (c) number of newly built residential community each year.

3.2. Results

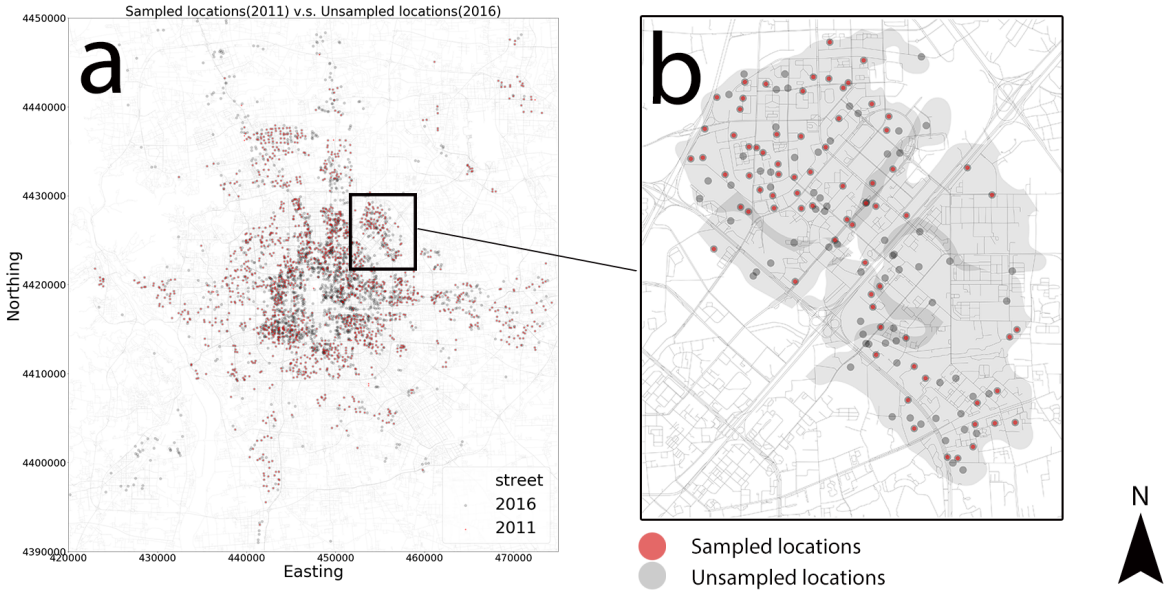

Figure 3 Spatial distributions of house prices at 2011 and 2016 (a) sampled locations in 2011 and Unsampled locations in 2016. (b) all communities within the area of *Wangjing, Beijing*.

We aggregated the house price data within the same year into the community-level, and formalized the data as spatiotemporal points (geometric centers of communities) on the map. The problem is simplified as "*we only have the sampled price for communities that have transaction records in 2011. Assuming it's now the end of 2015, there are some newly opened communities together with previous communities that need the price estimation for 2016*". The data is organized according to Figure 3.

To make it simple in this abstract, we selected *Wangjing* (a central business district in Beijing) as our target region (indicated in Figure 3b) and estimated its price distribution in 2016. A preliminary result is plotted in Figure 4. We constructed a spatial weighted graph based on the Euclidean distance using Equation (3) and α=4e2 (Figure 4a). The result is compared with the inverse distance weighted interpolation (IDW) with the same distance decay function. Noting that covariates such as the surrounding configuration and renovation condition may influence house price greater than the location, our graph is expected to be refined in future works to achieve better results.

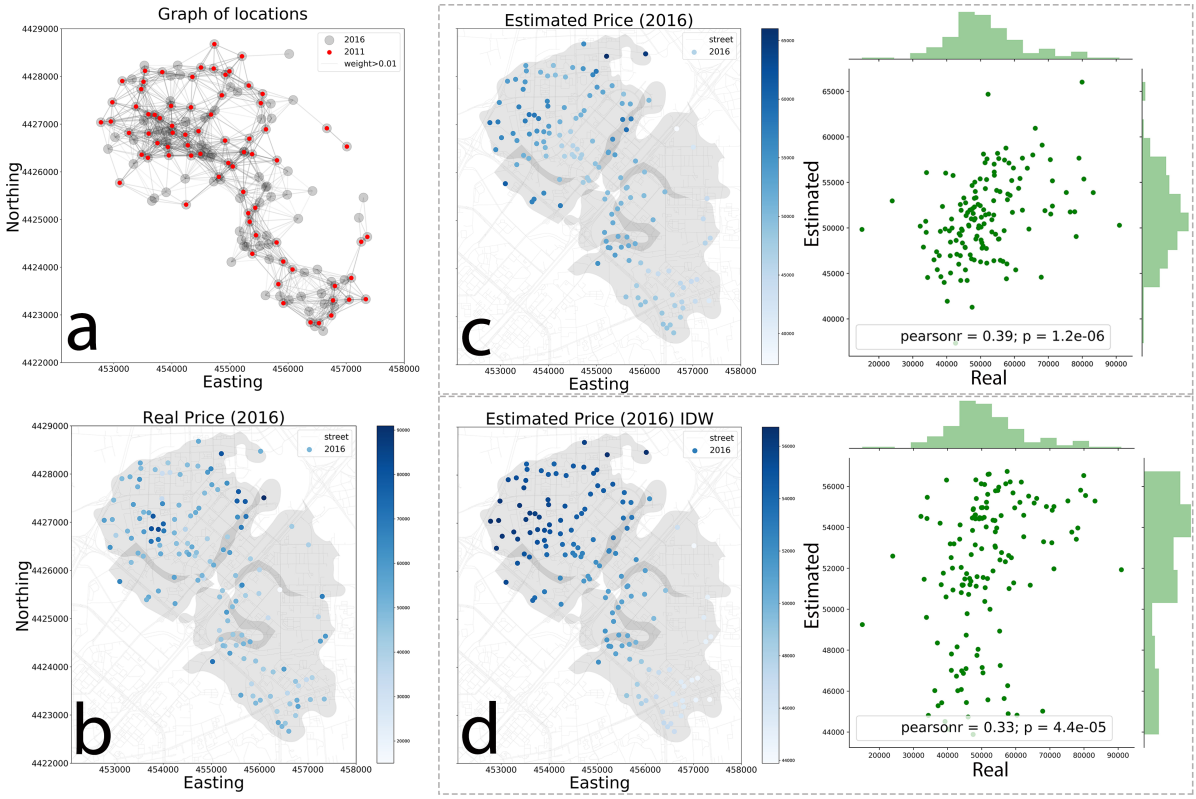

Figure 4 Preliminary result for 2016 (a) visualization of the graph. (b) real price distribution. (c) and (d) results of geo-propagation and IDW, accordingly.

4. Conclusion and Future lines

This paper proposes an idea named geo-propagation that formalizes the relations of locations into graphs and uses semi-supervised learning to achieve spatial estimation. A case study of sparse recorded house price is presented to show the feasibility of our geo-propagation method. Future works need to consider problems such as MAUP and ecology fallacy in the data aggregation. Temporal information needs to be better modelled to deal with the temporal scarcity and variation. Also, artificial intelligent algorithms that support graph representation learning are expected to be integrated into geo-propagation.

5. Acknowledgements

This research was supported by the China Scholarship Council Funding (No. 20180601077).

References

Atkinson, P.M. and Lloyd, C.D., 2009. Geostatistics and spatial interpolation. *The SAGE Handbook of Spatial Analysis*, 159-181.

Goodchild, M.F., Anselin, L., and Deichmann, U., 1993. A framework for the areal interpolation of socioeconomic data. *Environment & Planning A*, 25 (3), 383-397.

Goodchild, M.F., 2004. GIScience, Geography, Form, and Process. *Annals of the Association of American Geographers*, 94 (4), 709-714.

Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46, 234-240.
Zhu, D., Huang, Z., Shi, L., Wu, L., & Liu, Y. (2018). Inferring spatial interaction patterns from sequential snapshots of spatial distributions. International Journal of Geographical Information Science, 32(4), 783-805.
Zhu, A. X., Lu, G., Liu, J., Qin, C. Z., & Zhou, C. (2018). Spatial prediction based on Third Law of Geography. *Annals of GIS*, 1-16.
Zhu, X., Lafferty, J., & Rosenfeld, R. (2005). Semi-supervised learning with graphs (Doctoral dissertation, Carnegie Mellon University).

Biographies

Di Zhu is a visiting researcher at University College London and a PhD student at Peking University. His research interests include geospatial modelling, social sensing, economic geography and deep learning.

Tao Cheng is a professor in GeoInformatics at UCL. She is the Founder and Director of SpaceTimeLab for Big Data Analytics. Her research interests include network complexity, geocomputation, space-time analytics and big-data mining with applications in transport, crime, health, social media, and natural hazards.

Yu Liu is a professor in the Institute of Remote Sensing and Geographic Information System, Peking University. His research interests include humanities and social science based on big geo-data.