A Semi-supervised Regression Method Based on Geography Weighted Regression Model

Ruren Li[1], Hongming Wang[1], Guangchao Li[2](✉), and Haibo Qi[3]

[1] School of Transportation Engineering, Shenyang Jianzhu University,
Shenyang 100168, China
[2] College of Geoscience and Surveying Engineering,
China University of Mining and Technology, Beijing 100083, China
rurenli@163.com
[3] School of Mining and Geomatics Engineering,
Hebei University of Engineering, Handan 056038, China

Abstract. Aiming at the problem that the geometric weighted regression method has low prediction accuracy when the amount of training data is small, this paper combines the geometric weighted regression model with semi-supervised learning theory to make full use of semi-supervised learning that uses unlabeled samples to participate in training process to enhance the performance of the regression model, a semi-supervised regression model based on geo-weighted regression model is proposed, that is Self-GWR and CO-GWR. Based on the data of elevation, Aerosol Optical Depth (AOD), temperature, wind speed, humidity and pressure of Beijing-Tianjin-wing area, this paper uses geo-weighted regression model, self-trained geography weighted regression model and collaborative training geography weighted regression model respectively to practice. The experimental results show that CO-GWR effectively improves the accuracy of regression model through two regression models, and the accuracy of Self-GWR is slightly lower than that of GWR model, which indicates that the model may accumulate errors in the self-learning process, resulting in poor regression accuracy finally.

Keywords: GWR · Semi-supervised learning · Collaborative training · AOD

1 Introduction

In geospatial spatial analysis, the observed data is obtained in different geographical locations. The global spatial regression model assumes that the regression parameters are independent of the geographical location of the sample data, and in the actual problem research, it is often found that the regression parameters vary with the geographical location. If the global spatial regression model is still used, the regression parameter estimates obtained will be the average of the regression parameters in the whole study area, and cannot reflect the real spatial characteristics of the regression parameters [1, 2]. When spatial data has autocorrelation, the GWR (geographically weighted regression) model provides an estimation method superior to the ordinary linear regression model (OLS). The OLS method only provides estimation of global

Y. Xie et al. (Eds.): GSES 2019/GeoAI 2019, CCIS 1228, pp. 93–102, 2020.
https://doi.org/10.1007/978-981-15-6106-1_7

parameters, and the GWR model allows decomposition into local parts. The parameter estimation profoundly explains the relationship between certain types of indicators and spatial impact factors of geospatial data, which is unmatched by the OLS method [3]. The Geographically Weighted Regression (GWR) is a typical local model. The GWR model considers that the regression coefficient changes with the spatial position and can reflect the local relationship between the dependent variable and multiple independent variables [4]. Spatial non-stationary [5]. GWR modeling is performed using labeled samples. The accuracy of the GWR model is related to the number of labeled samples. The less the labeled samples, the lower the accuracy of the model. However, some labeled samples are difficult to obtain in large quantities, and unlabeled samples can be obtained in large quantities. Therefore, how to improve the accuracy of the GWR model with unlabeled samples is of great significance in the case of few labeled samples [6].

Semi-supervised learning samples with labeled and unlabeled samples can make use of a large number of unlabeled samples that are easily accessible, so that the workload of labeled samples is alleviated and a more efficient learning model is obtained [7]. Collaborative training uses two-view to train two classifiers to mark samples to each other to expand the training set. Using unlabeled samples to improve learning performance is an important method for semi-supervised learning [8]. Yang Y et al. [9] demonstrated that collaborative training using unlabeled sample-assisted training improves the learning performance when the labeled samples are small. The main direction of the current research is the clustering problem. The research on the regression problem is relatively rare, mainly because the clustering hypothesis in semi-supervised learning does not hold on the regression problem, and the confidence level is difficult to calculate in the regression analysis. In response to the above problems, the cooperative regression calculation method generates different k-nearest neighbor regression models based on different k-values or distance metrics, and selects unmarked samples with high confidence according to the prediction consistency for marking [10]. Zhao Yangyang [11] proposed a synergistic spatiotemporal geographic weighted regression method. The results show that the performance of spatiotemporal geo-weighted regression models with different kernel functions is not as high as that of geospatial-weighted regression models. Chai Yan [12] proposed a method for extracting boundary vectors based on the characteristics that support vectors are generally located at the boundary of two types of sample sets, and improved the SVM (support vector machines) algorithm. Ma and Wang et al. [13] proposed a regression model based on SVM cooperative training, which is suitable for the model when dealing with a large number of inputs, which alleviates the error accumulation problem caused by using only a single regression model, and improves the pan of the regression model. Ability. Brefeld et al. [14] proposed a semi-supervised least squares regression method to apply collaborative training to the normalized risk minimization problem in Hilbert space.

In summary, there are relatively few studies on semi-supervised regression learning methods, and there are fewer studies in the field of geographic information combined with semi-supervised regression learning. In this paper, for the problem of geographically weighted regression in the case of a small number of labeled samples, the prediction accuracy of the regression model is low. A model combining semi-supervised learning with geographically weighted regression is proposed, namely self-training geographically weighted regression model (Self-GWR) and collaborative training geography. This method makes full use of the advantages of semi-supervised learning, improves the regression accuracy of GWR under small sample data by unlabeled sample assistance, and can study the non-stable characteristics of space, which is more suitable for the analysis application of space field. The weighted regression model (CO-GWR) was used as the test data for the elevation, aerosol optical thickness, temperature, wind speed, humidity, and pressure in the Beijing-Tianjin-Wing region. By comparing with the conventional geographically weighted regression method, the mean absolute error (MAE), root mean square error (RMSE), Akaike information (AIC), and goodness of fit (R2) are used as evaluation indicators to verify the method. Effectiveness.

2 Methods

2.1 Geographically Weighted Regression

The geographically weighted regression was proposed by Fortheringham et al. [15] of the University of St. Andrews in the United Kingdom based on the regression of spatial coefficient of variation using the idea of local smoothness. Geographically weighted regression is an extension of ordinary linear regression, which introduces the geographic location of the sample points into the regression parameters. The formula is as follows:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{d} \beta_k(u_i, v_i)x_{ik} + \varepsilon_i \tag{1}$$

In Eq. (1), the n variables of the dependent variable y $(y_i, x_{i1}, x_{i2}, \cdots x_{id})$ and the independent variable $(x_1, x_2, \cdots x_d)$ at the data point (u_i, v_i), $(k = 1, 2, \cdots, d)\beta_k(u_i, v_i)$ are the unknown parameters at the observation point (u_i, v_i). For independent and identically distributed error terms, it is usually assumed to be subject to $N(0, \delta^2)$ distribution.

The regression parameters in the geographic weighted regression model (GWR) are related to the geographical location of the sample data, and the degree of influence (space weight) can be represented by a distance function, which is referred to as the kernel function. Commonly used kernel functions are Gaussian kernel functions, Bisquare kernel functions, and so on. The key to the GWR model is to choose the kernel function and determine its optimal bandwidth. The study found that the bandwidth sensitivity of different kernel functions is different, and the change of bandwidth will have a greater impact on the results [6]. Therefore, the regression model can be distinguished by kernel function and bandwidth.

If the bandwidth of the kernel function is too large, the regression parameter estimation is too large, and too small will cause the regression parameter estimation to be too small [1]. In order to reduce the error caused by bandwidth discomfort, this paper uses Clevel cross-validation method [16] proposed by Cleveland to calculate the optimal bandwidth. The CV method is calculated as:

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \widehat{y_{\neq i}}(b)\right]^2 \tag{2}$$

In formula (2), when the regression parameter $\widehat{y_{\neq i}}(b)$ is estimated, the regression point itself is not included. Only the regression parameter calculation is performed according to the data around the regression point, and the different bandwidth and CV are drawn into the trend line, and the minimum CV(b) value can be found and Its corresponding optimal bandwidth is b.

2.2 Semi-supervised Learning

Self-training Geographically Weighted Regression. In semi-supervised learning, Fralick et al. [17] proposed a self-training learning method. In each round of training, the best sample from the previous round of predictions is added to the current set of labeled samples. The results produced continue to train themselves [18]. Based on the geographically weighted regression model and the semi-supervised self-training theory, this paper proposes a self-training geographic weighted regression model (Self-GWR).

The steps of the self-training geo-weighted regression model algorithm are as follows: 1 determining the labeled samples, unlabeled samples, initializing the GWR model parameters, using the Gaussian kernel function in the GWR model; 2 using the GWR model to perform regression prediction on the unlabeled samples; The highest degree of data is added to the labeled sample set of the regression model and the sample is removed from the unlabeled sample; 4 iteratively trains the GWR model until the unlabeled sample is trained to a certain number.

Collaborative Training Geographically Weighted Regression. Collaborative training is based on a small number of labeled samples and a large number of unlabeled samples. Through continuous iteration, different learners learn from each other [19], which is a semi-supervised learning method. The core idea of collaborative training is: First, use a labeled sample to train a classifier on each view. Then, each classifier selects a number of high-confidence samples from unmarked samples for marking, and puts these new tags. The sample is added to the training set of another classifier so that the other party can update with these new tag samples. This process of "learning from each other and making progress together" is iteratively continued until the two classifiers are not changing or reaching a predetermined number of learning rounds [20].

Based on the theory of geographically weighted regression model and cooperative training regression algorithm [13], combined with semi-supervised learning collaborative training theory, a collaborative training geographic weighted regression model (CO-GWR) is proposed. The CO-GWR model not only integrates the characteristics of the GWR model in geographic applications, but also compensates for the few defects of the sample. The algorithm steps are as follows: 1. Determine the labeled samples, unlabeled samples, initialize the GWR model parameters h1, h2, and h1 is based on Gaussian. The GWR model of the kernel function, h2 is the GWR model based on the Bisquare kernel function; 2 using each regression model to predict the regression of the unlabeled sample set, and selecting the labeled sample set added to the regression model with the highest confidence in the prediction result. And the sample is removed from the unlabeled sample; 3 repeat the second step operation until the unlabeled sample is trained to a certain number; 4 finally take the average of the two model prediction results as the final prediction result.

Confidence. Confidence is the degree to which the unlabeled data in the regression model affects the accuracy of the regression model during training. The higher the confidence, the better the consistency of the prediction and regression models, and the closer to the true value. Therefore, the sample selected by the regression model through high confidence should be a sample that makes the regression model more consistent with the labeled sample [20]. Confidence is based on MAE and its calculation method is:

$$\xi X_{x\int u} = \sum_{x_L} \left[(y_L - \hat{y}_L)^2 - (y_L - \hat{y}'_L)^2 \right] \tag{3}$$

In Eq. (3), to mark the true value of the sample, y_L is the predicted value of the labeled sample on the original regression model, y_L is the predicted value of the labeled sample on the new regression model, and the new regression model refers to adding the unlabeled Regression model re-established after the sample. At the $\xi X_{x \in u} > 0$ time, order. $N(x, u, v) = arc\max(\xi X_{x \in u})$. The most unmarked sample $N(x, u, v)$ with the highest confidence. $\xi X_{x \in u} > 0$ shows that the performance of the regression model is improved after the unlabeled samples are added. The maximum confidence indicates that the performance of the model is the largest, that is, the selected sample is the one with the highest confidence among the unlabeled samples participating in the training.

3 Overview of the Study Area

The study area uses the Beijing-Tianjin-Hebei region with a geographical range of 35.5° N −43°N and 113°E −120°E, including 11 prefecture-level cities in Beijing, Tianjin and Hebei Province. It is located in the heart of China's Bohai Sea and is China. The largest and most dynamic region in the north has a land area of more than 200,000 and a total population of more than 100 million. With the rapid development of the economy, air pollution is becoming more and more serious, and Beijing-Tianjin-Hebei is also one of the heavily polluted areas. Therefore, controlling air pollution and optimizing the ecological environment are important tasks in the Beijing-Tianjin-Hebei region. The changes in atmospheric aerosol types and contents are closely related to climate change and atmospheric environmental pollution. The research on AOD is of great significance for the analysis and prevention of atmospheric environmental pollution.

4 Application to Simulated Data

4.1 Data Source

This paper takes geospatial data (elevation), meteorological data (wind speed, temperature, humidity, air pressure) and AOD data in the Beijing-Tianjin-Hebei region as research objects. The geospatial data comes from the Geospatial Data Cloud website (http://www.gscloud.cn/), and the SRTMEDM 90 M resolution raw elevation data is selected; the meteorological data is from the China Meteorological Science Data Sharing Service Network (http://www.escience.gov.cn) A total of 110 meteorological monitoring sites with geographic location information for one day; AOD data from the Terra MODIS C06 secondary aerosol product, the frequency is one day, and the spatial resolution is 3 km. In this paper, MODIS Collection 6 MYD04_3K data set parameter named "Optical_ Depth_ Land_ And_ Ocean", band 2 550 nm Class 2 AOD data, AOD data selected May 2015 data as the research object. The meteorological monitoring stations were randomly divided into marked monitoring stations and unmarked monitoring stations in a 1:1 ratio (Fig. 1).

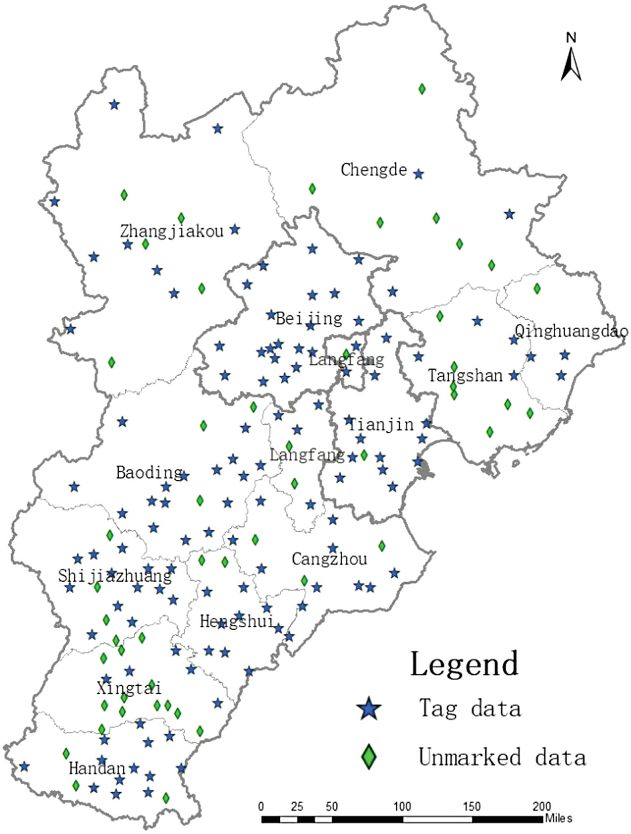

Fig. 1. Meteorological monitoring site distribution

4.2 Data Preprocessing

For the spatial and temporal consistency of the data, the geospatial data, meteorological data, and AOD data are preprocessed. The preprocessed partial data is shown in Table 1, and the preprocessing flowchart is shown in Fig. 2. Firstly, the meteorological data is processed to obtain the daily mean value of 110 meteorological data and its monthly mean value. Secondly, the geospatial data is processed, including the projection coordinate transformation of the image data, and the spatial data value of the meteorological site for the elevation image data. Finally, for AOD data processing, a 5 km × 5 km grid covering the whole area is created for the Beijing-Tianjin-Hebei region, and the position coordinates of the grid center point are extracted, and the grid center point represents the spatial position of the grid. The MODIS image data is processed in batches, the value of the AOD of the grid center point is obtained by resampling, and the monthly mean value of the AOD data of the grid center point is calculated, and the monthly mean value is interpolated by Kriging to extract the value of the AOD at the weather station.

Table 1. Partial data display

ID	Site number	Latitude	Longitude	Elevation	Average wind	Average temperature	Average humidity	Air pressure	AOD
1	53392	41.85	114.6	1418	2.1167	17.1833	26.6667	853.4	0.5282
2	53893	36.7667	114.95	42	1.8885	23.6167	53.25	1003.9296	0.5377
3	54809	36.55	115.2833	42	1.7167	21.9833	56.3333	1004.2833	0.5709
4	53899	36.4833	114.9667	48	2.1417	22.1083	58.3333	1003.7167	0.4822
5	53693	38.0333	114.1333	215	2.8417	23.975	48.3333	978.6333	0.4044
6	54412	40.7333	116.6333	285	1.75	14.2	84	971.95	0.3699
7	54541	39.9	119.2333	28	2.4	23.5083	39.6667	1005.6667	0.5824
8	54301	41.6667	115.6667	1412	2.175	17.45	32.6667	854.7917	0.0865

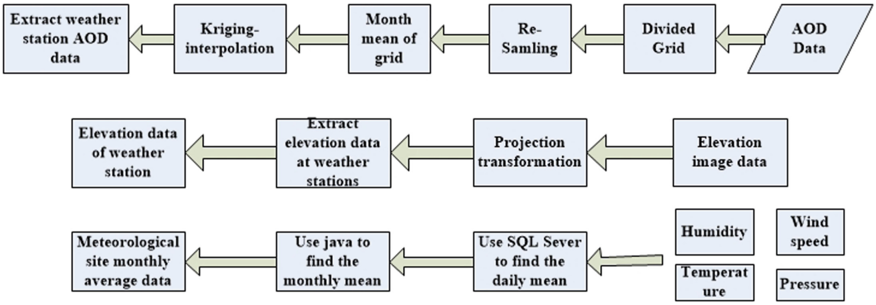
Extract weather station AOD data
Kriging-interpolation
Month mean of grid
Re-Samling
Divided Grid
AOD Data
Elevation data of weather station
Extract elevation data at weather stations
Projection transformation
Elevation image data
Meteorological site monthly average data
Use java to find the monthly mean
Use SQL Sever to find the daily mean
Humidity
Wind speed
Temperat ure
Pressure

Fig. 2. Data preprocessing flow chart

4.3 Results and Analysis

The correlation coefficient between each index is calculated, and the correlation coefficient matrix is shown in Table 2. It can be seen from Table 2 that the five indicators considered in this paper are related to AOD to a certain extent. Humidity, pressure and temperature are positively correlated with AOD, and elevation and wind speed are negatively correlated with AOD, including elevation, temperature and pressure. The correlation coefficient with AOD is more than 0.62. (The following data retains 4 significant figures)

Table 2. Correlation coefficient matrix

	AOD	Elevation	Air temperature	Humidity	Wind speed	Air pressure
AOD	1	−0.6238	0.6454	0.3951	−0.2053	0.6225
Elevation		1	−0.9398	−0.7152	0.5270	0.9921
Air temperature			1	0.5347	−0.4799	0.9388
Humidity				1	−0.5259	0.7247
Wind speed					1	−0.5373
Air pressure						1

In order to evaluate the prediction effect of the research method, this paper compares it with the GWR model, and calculates the four evaluation indexes of MAE, RMSE, AIC and R2 of each model, and the degree of improvement between the models. MAE reflects the possible error range of the estimated value, RMSE reflects the inversion sensitivity and extremum effect of the interpolation function, and the AIC criterion is based on the concept of entropy, which can weigh the complexity of the estimated model and the superiority of the model fitting data. R2 can measure the pros and cons of the model fit. The calculation results are shown in Table 3.

Table 3. Comparison of forecast results

Mode	MAE	RMSE	AIC	R^2
Self-GWR	0.3604	0.4497	162.2593	0.6513
CO-GWR	0.3137	0.3971	143.3187	0.7386
Self-GWR/GWR upgrade	−4.71%	−4.29%	−5.06%	−5.91%
CO-GWR/GWR upgrade	8.86%	7.91%	7.20%	6.70%

5 Conclusions

From Table 2, we can see that the GWR model is 0.3442, 0.4312, 154.4414 and 0.6922, the Self-GWR model is 0.3604, 0.4497, 162.2593 and 0.6513, and the CO-GWR model is 0.3137, 0.3971, 143.3187 and 0.7386, respectively, among the evaluation indexes MAE, RMSE, AIC and R2 obtained by each model. Compared with

GWR model, CO-GWR model increased percentages by 8.86, 7.91, 7.20 and 6.70, respectively. The experimental results show that the accuracy of the Self-GWR model is the worst among the three models, mainly because the model in the self-training process, the error occurred in the previous stage will accumulate and amplify in the subsequent learning process, so that the model final accuracy is poor. The CO-GWR model has the highest precision and the best effect, which fully demonstrates that the collaborative training method not only plays the role of unlabeled samples in semi-supervised learning, but also alleviates the deficiencies of self-training model error accumulation and improves the generalization performance of the model. The prediction accuracy of the regression model is improved.

This paper combines semi-supervised learning with a geographically weighted regression model, namely self-training geographic weighted regression model (Self-GWR) and collaborative training geographic weighted regression model (CO-GWR). Both models can use unlabeled samples to improve the performance of the regression model, and can also analyze the non-stationary features of the geographical phenomenon, so that the semi-supervised learning method can be better applied in the field of spatial analysis. By comparing with the conventional GWR model, it shows that the CO-GWR model effectively improves the accuracy of the regression model by means of two regression models “mutual learning”, while the accuracy of the Self-GWR model is slightly lower than that of the GWR model, indicating that the model is Accumulation of errors may occur during self-learning, resulting in poor final regression accuracy. In the follow-up study, the above research methods will continue to be improved. For example, for collaborative training, two different regression models can be used for analysis. For self-training, outliers in the data can be eliminated to achieve better learning results.

References

1. Yan, W.: Basic Theory and Application of Geographically Weighted Regression. Tongji University, Shanghai (2007)
2. Yang, Y.: Study on the Non-stationary Geographically Weighted Regression Method of Time and Space. Wuhan University, Wuhan (2016)
3. Tang, Q., Xu, W., Ai, W.: Study on spatial differentiation of house price and its influencing factors based on geographically weighted regression. Econ. Geogr. (02), 52–58 (2012)
4. Ren, Q, Wang, L., Li, H.: Analysis of spatial differences of regional economic development in China. Geogr. Geogr. Inf. (01), 110–116 (2017)
5. Wu, B., Barry, M.: Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. pp. 383–401. Taylor & Francis Inc. (2010)
6. Zhao, Y., Liu, J., Xu, S., et al.: A geographically weighted regression method based on semi-supervised learning. J. Surv. Mapp. (01), 123–129 (2017)
7. Ma, L., Wang, X.: Semi-supervised regression based on cooperative vector machine cooperative training. Comput. Eng. Appl. (03), 177–180 (2011)
8. Guo, X., Wang, W.: An improved collaborative training algorithm: compatible co-training. J. Nanjing Univ. (Nat. Sci.) (04), 662–671 (2016)

9. Yang, Y., Liu, J., Xu, S., et al.: An extended semi-supervised regression approach with co-training and geographical weighted regression: a case study of housing prices in Beijing. ISPRS Int. J. Geo-Inf. **5**(1), 4 (2016)
10. Zhou, Z.H., Li, M.: Semi-supervised regression with co-training. In: International Joint Conference on Artificial Intelligence (2005)
11. Zhao, Y., Liu J., Yang, Y., et al.: A method for estimating PM2.5 concentration in synergistic space-time geographically weighted regression. J. Surv. Mapp. (12), 172–178 (2016)
12. Chai, Y., Wang, Y., Zhang, J.: Support vector machine algorithm under boundary vector. J. Liaoning Techn. Univ. (Nat. Sci. Ed.) (02), 202–205 (2017)
13. Lei, M.A., Wang, X.: Semi-supervised regression based on support vector machine co-training. Comput. Eng. Appl. **25**(2) (2011)
14. Efron, B., Gong, G.: A leisurely look at the bootstrap, the jackknife, and cross-validation. Am. Stat. (1), 36–48 (1983)
15. Fotheringham, A.S., Charlton, M., Brunsdon, C.: Measuring Spatial Variations in Relationships with Geographically Weighted Regression. Springer, Heidelberg (1997)
16. Cleveland, W.S.: Robust locally weighted and smoothing scatterplots. J. Am. Stat. Assoc. **74** (368), 829–836 (1979)
17. Fralick, S.: Learning to recognize patterns without a teacher. IEEE Trans. Inf. Theory **13**(1), 57–64 (2003)
18. Liu, J., Liu, Y., Luo, X.: Semi-supervised Learning Method. Chin. J. Comput. **08**, 1592–1617 (2015)
19. Zhou, Z.H., Li, M.: Semisupervised regression with cotraining-style algorithms. IEEE Trans. Knowl. Data Eng. **19**(11), 908–913 (2007)
20. Zhou, Z.: Semi-supervised learning based on bifurcation. Acta Autom. Sinica **11**, 1871–1878 (2013)