GWRBoost:
A geographically weighted gradient boosting method for explainable quatifications of spatially varying relationships

Han Wang · Institute of Remote Sensing and GIS
HanwGeek@gmail.com

北京大學
PEKING UNIVERSITY

Contents

PRELIMINARIES

Linear Regression

Classical linear regression for N independent observations with K variables:

dependent variable

estimated coefficients

$$\mathrm{y_i} = \beta_0 + \sum_{\mathrm{k}=1}^{\mathrm{K}} \beta_{\mathrm{k}} \mathrm{x_{ik}} + \varepsilon \tag{1}$$

independent variables

random error term $\sim \mathcal{N}(\mu, \sigma^2)$

Least square method to estimate coefficients:

$$\min_{\boldsymbol{\beta}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \left[\mathrm{y_i} - \left(\beta_0 + \sum_{\mathrm{k}=1}^{\mathrm{K}} \beta_{\mathrm{k}} \mathrm{x_{ik}}\right) \right]^2 \tag{2}$$

Coefficients β_{k} represents the explainable quantification of the relationship between $\mathrm{x_k}$ and y.

Geographically weighted regression (GWR)

[Spatial heterogeneity] In a geographical context,

relationships between variables may not be constant. (Goodchild, 2004)

[Solution] Localized coefficients

coefficients specifically for location (u_i, v_i)

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{K} \beta_k(u_i, v_i)\, x_{ik} + \varepsilon \tag{3}$$

with weighted least square for i-th observation

$$\min_{\boldsymbol{\beta}(u_i, v_i)} \sum_{i=n}^{N} W\left[(u_i, v_i), (u_n, v_n)\right] \cdot \left[y_n - \left(\beta_0(u_i, v_i) + \sum_{k=1}^{K} \beta_k(u_i, v_i)\, x_{nk}\right)\right]^2 \tag{4}$$

spatial weight for the observation at location (u_n, v_n)

Weights are generated by spatial kernel functions (e.g. bi-square)
Generate a better estimates than oridinary least square(OLS)

Issues

1. Linear models are prone to be underfitting (Schell & Singh, 1997)

- inferior performance compared with more complex models (e.g. decision tree, SVM, neural network)

[Solution]

- Geographically neural network weighted regression (Du et al., 2020)
- Spatial regression graph convolutional neural networks (Zhu et al., 2021)

However,

- sophisticated models cannot generate explicit coefficients for relationships quantification [XGBoost + SHAP (Li, 2022)]
- hard to be evaluated via AIC score fairly (risk of being overfitting)

Issues

2. Weighted least square may not achieve global optimal

- different from the global evaluation metric (e.g. AIC, RSS)
- each observation optimizes the local parameters independently

Develop a model that

- has higher model complexity to handle large volume data
- can generate explainable quantification of spatially varying relationships
- maintain the global optimization objective function (consistent with the evaluation metrics)

GWRBoost

Ensemble learning & Boosting algorithm

[step-wise optimization] Gradient descent

General model parameters

$$\theta_i = \theta_{i-1} + \Delta\theta_i = \theta_{i-1} - \lambda \left[\underbrace{\frac{\partial \mathcal{L}(\theta)}{\partial \theta}}_{\text{gradient of } \theta} \right]_{\theta=\theta_{i-1}} \quad (5)$$

Global non-parametric model $F_\theta(x)$

$$F_i = F_{i-1} + \Delta F_i = F(x)_{i-1} - \lambda \left[\underbrace{\frac{\partial \mathcal{L}(F)}{\partial F}}_{\text{gradient of } F} \right]_{F=F_{i-1}} \quad (6)$$

[Additive model]

Base machine learning model

$$f(x) \sim -\lambda \left[\frac{\partial \mathcal{L}(y, F(x))}{\partial F(x)} \right]_{F=F_{i-1}(x)} \Rightarrow \sum f(x) = F(x) \quad (7)$$

Ensemble learning

Accumulate simple & weak base models to

- increase the model capacity
- generate better results

[Local] Localized additive model

Linear regression

Still linear form

$$y_i = F_i(x_i) = \sum_{m=1}^{M} f(x;\beta^m) = \sum_{m=1}^{M} \beta_0^m + \sum_{m=1}^{M}\sum_{k=1}^{K} \beta_k^m (u_i, v_i)\, x_{ik} + \varepsilon_i \tag{8}$$

[Global]

$$F(x) = \{F_1, F_2, \dots, F_N\} \tag{9}$$

[Gradient boosting optimization]

Residuals from the previous step

$$f_{\beta^m}(x_i) \sim \lambda \frac{\partial \mathcal{L}}{\partial F^{m-1}} = \lambda \frac{\partial \left[y - F^{m-1}(x)\right]^2}{\partial F^{m-1}} = \lambda \cdot \frac{1}{N} \sum_{i=1}^{N} \left[y_i - F^{m-1}(x_i) \right] \tag{10}$$

Geographically weighted regression

Learn the residuals instead of ground truth in a geographically weighted way

GWRBoost

Algorithm 1: GWRBoost

Data: $D = \{(X_1, y_1, u_1, v_1), \ldots, (X_N, y_N, u_N, v_N)\}$
Result: Model set $\mathcal{F}^{\mathcal{M}} = \{F_1^M, \ldots, F_N^M\}$

```
for n = 1 to N do
    β̂_n^1 = arg min_{β_n^1} (1/2) Σ_{i=1}^N w_i (y_i − f_{β_n^1}(x_i))^2
    F_i^1 = f_{β_n^1}
end
for m = 2 to M do
    for n = 1 to N do
        r_n = λ · [y_n − F_n^{m−1}(x_n)]
        β̂_n^m = arg min_{β_n^m} (1/2) Σ_{i=1}^N w_i (r_n − f_{β_n^m}(x_i))^2
        F_n^m = F_n^{m−1} + f_{β_n^m}
    end
end
```

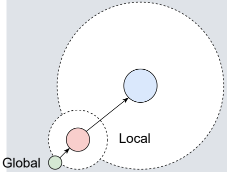
$$\hat{\beta}_n^1 = \arg\min_{\beta_n^1} \tfrac{1}{2} \sum_{i=1}^{N} w_i (y_i - f_{\beta_n^1}(x_i))^2, \quad F_i^1 = f_{\beta_n^1}$$

$$r_n = \lambda \cdot [y_n - F_n^{m-1}(x_n)], \quad \hat{\beta}_n^m = \arg\min_{\beta_n^m} \tfrac{1}{2} \sum_{i=1}^{N} w_i (r_n - f_{\beta_n^m}(x_i))^2, \quad F_n^m = F_n^{m-1} + f_{\beta_n^m}$$

Summary

- Initialize a GWR
- Collect the residuals
- new GWR models are trained to fit the residuals continuously

Residual passing



Capture global information by collecting residuals gradually

Computation of AIC

likelihood function

degree of freedom: trace of hat matrix $\mathcal{H}$

$$\mathrm{AIC} = -2\ln(\hat{\mathcal{L}}) + 2\mathrm{k} \tag{11}$$

mapping from y to $\hat{\mathrm{y}}$

$$\hat{\mathrm{y}} = \mathrm{X}\hat{\beta} = \mathrm{X}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y} = \mathcal{H}\mathrm{y} \tag{12}$$

$$\hat{\mathrm{y}} = \sum_{\mathrm{m}=1}^{\mathrm{M}}\hat{\mathrm{y}}_{\mathrm{m}} = \sum_{\mathrm{m}=1}^{\mathrm{M}}\mathcal{H}\mathrm{y}_{\mathrm{m}} = \mathcal{H}\sum_{\mathrm{m}=1}^{\mathrm{M}}(\mathrm{I}-\mathcal{H})^{\mathrm{m}-1}\mathrm{y}_1 = \left\{\mathcal{H}\sum_{\mathrm{m}=1}^{\mathrm{M}}[\lambda(\mathrm{I}-\mathcal{H})]^{\mathrm{m}-1}\right\}\mathrm{y} \tag{13}$$

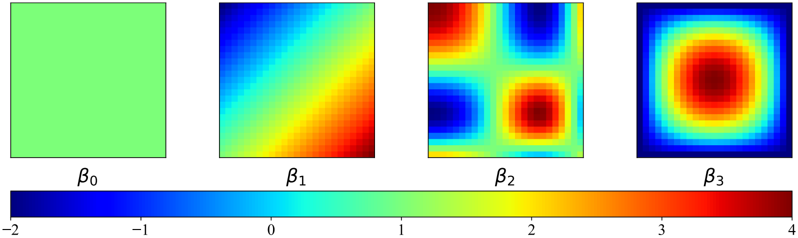
hat matrix of M learners

AIC of complex models (e.g. decision tree, neural network) is hard to measure.

EXPERIMENTAL EVALUATION

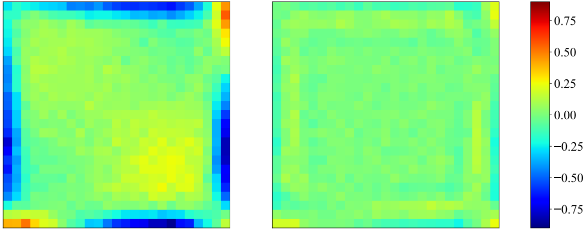
Synthetic data

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{14}$$



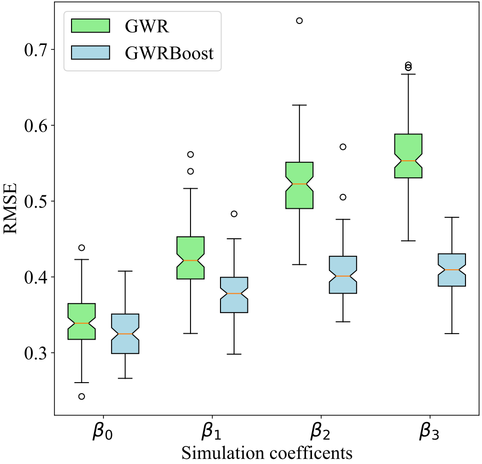
Model	OLS	GWR	**GWRBoost**
RSS	1639.063 ± 72.52	83.900 ± 5.049	**36.797** ± **2.601**
AIC	2385.642 ± 27.65	773.374 ± 36.050	**225.512** ± **42.061**
AICc	2385.739 ± 27.65	839.926 ± 35.383	**274.817** ± **41.207**
R^2	0.072 ± 0.02	0.952 ± 0.003	**0.979** ± **0.002**
Adjusted R^2	0.066 ± 0.02	0.940 ± 0.004	**0.975** ± **0.002**

Errors of coefficients estimates



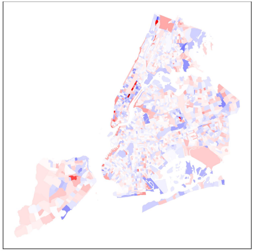
(a) GWR (b) GWRBoost

Figure: The distribution of residuals

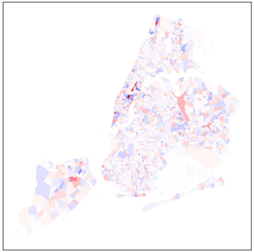


- reduction of marginal fitting errors
- lower RMSE of coefficients estimation
- lower variance in error bound
- higher performance in stronger relationships

Empirical case study — NYC education data



Spatial distritbution of GWR residuals



Spatial distritbution of GWRBoost residuals

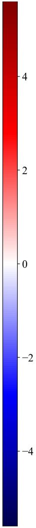


Table: Selected indicators

Variable	Explanation
Dependent	
mean_inc	Mean income
Independent	
sub18	Population under 18 years (count)
PER_PRV_SC	Percentage of all students enrolled in private school
YOUTH_DROP	Percentage of population age 16-19 that has dropped out of high school
HS_DROP	Percentage of population age over 25 that dropped out of high school
COL_DEGREE	Percentage of population age over 25 that obtained at least a bachelors degree
SCHOOL_CT	Number of schools (count)

Table: Comparative performance

Model	OLS	GWR	**GWRBoost**
RSS	982.206	388.626	**261.478**
AIC	4499.669	3168.118	**2289.994**
AICc	4499.720	3315.637	**2437.513**
R^2	0.557	0.825	**0.882**
Adjusted R^2	0.556	0.790	**0.858**
Moran’s I	0.333	0.066	**-0.027**

CONCLUSION

Conclusion

How it works? Why it works?

- learn residuals instead of ground truth to maintain appropriate objectives
- collect global information by residual passing process

Conclusion: We proposed a framework that

- applies gradient boosting algorithm for optimization
- increase the model complexity to maintain large volume data
- can be evaluated by AIC/AICc
- retains the ability to generate explainable & explicit quantifications for spatially varying relationships

Further issues

[Computational overhead] · [Bandwidth selection] · [More ensemble learning]

Thanks!

[Preprint] Please see this:

https://arxiv.org/abs/2212.05814

[Contacts] Feel free to contact me!

- Email: hanwgeek@gmail.com
- Twitter: @HanwGeek
- LinkedIn: @HanwGeek

Q & A

References

- Du, Z., et al., 2020. Geographically neural network weighted regression for the accurate estimation of spatial non-stationarity. International Journal of Geographical Information Science, 34 (7), 13531377.
- Goodchild, M.F., 2004. The validity and usefulness of laws in geographic information science and geography. Annals of the Association of American Geographers, 94 (2), 300303.
- Li, Z., 2022. Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost. Computers, Environment and Urban Systems, 96, 10184.
- Schell, M.J. and Singh, B., 1997. The reduced monotonic regression method. Journal of the American Statistical Association, 92 (437), 128135.
- Zhu, D., et al., 2021. Spatial regression graph convolutional neural networks: A deep learning paradigm for spatial multivariate distributions. GeoInformatica, 132.