Project Information

Title: Generalized Modeling and Predictions in Multiscale Geographically Weighted Regression

Abstract:

A recent addition to the local statistical models in PySAL is the implementation of Multiscale Geographically Weighted Regression (MGWR) model, a multiscale extension to the widely used approach for modeling process spatial heterogeneity - Geographically Weighted Regression (GWR). GWR is a local spatial multivariate statistical modeling technique embedded within the regression framework that is calibrated and estimates covariate parameters at each location using borrowed data from neighboring observations. The extent of neighboring observations used for calibration is interpreted as the indicator of scale for the spatial processes and is assumed to be constant across covariates in GWR. MGWR, using a back-fitting algorithm relaxes the assumption that all processes being modeled operate at the same spatial scale and estimates a unique indicator of scale for each process.

The GWR model in PySAL can currently estimate Gaussian, Poisson and Logistic models though the MGWR model is currently limited to only Gaussian models. This project aims to expand the MGWR model to nonlinear local spatial regression modeling techniques where the response outcomes may be discrete (following a Poisson distribution) or binary (Logistic models). This will enable a richer and holistic local statistical modeling framework to model multi-scale process heterogeneity for the open source community. In addition, to support efficient testing for different model implementations, a simulated data generator module will be implemented to supply test datasets following unique model variable distribution needs. This will also provide a foundation for possible expansion to test other local model implementations in PySAL. GWR has been widely used as a tool for spatial prediction and has been known to be informative on the spatial processes generating the data being predicted (Harris et al., 2010). While the GWR implementation in PySAL facilitates the predictions for the dependent variable at unsampled locations, this functionality has not been implemented for MGWR yet. This project aims to also enable the prediction functionality for MGWR and solve for its growing need in the open source community (https://github.com/pysal/mgwr/issues/51). In doing so, open issues around predictions in GWR (for e.g. https://github.com/pysal/mgwr/issues/50) will also be resolved.

Expected theoretical model calibration:

The calibration methodology for modeling response variables with a Poisson distribution in MGWR, through references from Geographically Weighted Poisson Regression (GWPR) (Nakaya, et al., 2005) and literature on Generalized Additive Models (Hastie & Tibshirani, 1986), is expected to be as follows:

$O_i \sim Poisson[E_i exp(\sum_{k=1}^{K} \beta_k x_{i,k})]$ (Conventional Poisson Regression model)

$\widehat{O}_j(\beta(u_i)) = E_j exp(\sum_{k=1}^{K} \beta_k(u_i) x_{j,k})]$ (Geographically varying parameters in Poisson Regression),

where $\boldsymbol{X}$ is the (N, K) matrix of predictor variables and $\boldsymbol{W_i}$ is the (N, N) diagonal spatial weighting matrix for location i with the diagonal elements representing the weights attached to each location and is calculated based on a specified kernel function and bandwidth.

(1) Initialize using GWPR estimates $\beta_k(u_i)^0$: $\mathrm{f}_1^0, \mathrm{f}_2^0, \ldots, \mathrm{f}_K^0$, where $\mathrm{f}_k^0 = \begin{pmatrix} x_{1k}\beta_k(u_1)^0 \\ \ldots \\ x_{Nk}\beta_k(u_N)^0 \end{pmatrix}$

(2) Update for each location (u_i) an adjusted dependent variable

$$\mathrm{z}(u_i)^{(l)} = (\mathrm{z}_1(u_i)^{(l)}, \quad \mathrm{z}_2(u_i)^{(l)}, \quad \ldots \quad , \mathrm{z}_N(u_i)^{(l)})^t$$

$$\mathrm{z}_j(u_i)^{(l)} = \sum_{k=1}^{K} \beta_k(u_i)^{(l)} x_{j,k} + \frac{\mathrm{O}_j - \widehat{\mathrm{O}}_j\left(\beta(u_i)^{(l)}\right)}{\widehat{\mathrm{O}}_j(\beta(u_i)^{(l)})}$$

(3) Construct weights as follows:

$$\mathrm{A}(u_i)^{(l)} = \begin{pmatrix} \widehat{\mathrm{O}}_1\left(\beta(u_i)^{(l)}\right) & & 0 \\ & \widehat{\mathrm{O}}_2\left(\beta(u_i)^{(l)}\right) & \\ 0 & & \ddots \\ & & \widehat{\mathrm{O}}_N\left(\beta(u_i)^{(l)}\right) \end{pmatrix}$$

(4) Fit an MGWR model to $\mathrm{z}(u_i)^{(l)}$ to update $\beta_k(u_i)$ and $\mathrm{z}(u_i)^{(l)}$ using the new weight matrix:

$$W_k{}^*(u_i)^{(l)} = W_k(u_i)^{(l)}\mathrm{A}(u_i)^{(l)}$$

(5) Repeat steps (2) through (4) until convergence.

The calibration for binary response variables is expected to follow a similar iteration process with variations in values used for initializing and the link function.

Major Deliverables:

Core modules implementing generalized model calibration algorithms – one each for Poisson and Logistic regression extensions to MGWR models

Diagnostics modules for implementing model diagnostics (AIC, BIC, CV) - one for each model

Data simulation module for creating model specific simulated data for testing the models (can be extended to incorporate testing of existing MGWR and GWR models)

Memory and time Optimization enhancements for generalized MGWR model calibration and convergence (since the iterations in the calibration above include a backfitting algorithm within, the calibration is expected to be time and memory intensive)

MGWR Prediction module to implement predictions for the dependent variable in an MGWR model at unsampled locations

Empirical example studies with real datasets as a part of model helper notebooks to show implementation and results generated by the models (Poisson and Logistic) and prediction functionality

Proposal Timeline:

Before May 06:

- Understand and get familiarized with PySAL's 'model' module architecture and dependencies
- Extensive research on Generalized Additive Models for foundational understanding

May 06 – May 27

- Discuss scope and possible pre-requisites that may be required for the modules with mentor
- Finalize statistical approach hypothesized in proposal with input from mentor
- Discuss scope of diagnostics and inference for generalized MGWR models

May 27 – June 17

- Implement finalized calibration algorithm for Poisson response MGWR model
- Observe and tabulate convergence thresholds and behavior of iterations
- Observe and tabulate implementation time and system memory usage for iterations until convergence

June 17 – July 08

- Fix theoretical changes if any after reviewing results with mentor
- Introduce parameterized convergence thresholds based on observed behavior
- Study options for optimization enhancements for iterations

It is expected that implementation time and memory requirement will be high given the complex iteration form. This step will hence inform an important aspect of the model functionality.

July 08 – July 15

- Implement optimization algorithms from finalized options to reduce time and memory usage
- Write simulated data generator code for testing the model
- Write tests for the models utilizing data from data simulation module

July 15 – July 22

- Implement similar module for Logistic model for MGWR (since the calibration algorithms are expected to be similar to the one implement for Poisson MGWR, with minor initialization and link function tweaks, this is expected to take lesser time.)
- Implement optimization algorithms to optimize model calibration similar to that done for Poisson MGWR

July 22 – August 05

- Implement prediction functionality for MGWR
- Tests and observation of results with multiple datasets to ensure prediction accuracy

August 05 – August 19

- Documentation
- Example notebooks implementing the two models with real datasets

One buffer week for unanticipated delays or additions.

References:

Harris, P., Fotheringham, A. S., Crespo, R., & Charlton, M. (2010). The Use of Geographically Weighted Regression for Spatial Prediction: An Evaluation of Models Using Simulated Data Sets. *Mathematical Geosciences*, *42*(6), 657–680. https://doi.org/10.1007/s11004-010-9284-7

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, *1*(3), 297–310. https://doi.org/10.1214/ss/1177013604

Nakaya, T., Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine*, *24*(17), 2695–2717. https://doi.org/10.1002/sim.2129