Explaining Business Satisfaction: Identifying Local Factors with Geographic Weighted Regression

Douglas Callaway

October 2016

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Department of Computer Science

Supervisor: Professor Mark Gahegan

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Professional Studies in Data Science, The University of Auckland, 2016.

Abstract

Last

Contents

1. Introduction

Sixth

- Motive (why)
- Global model can obscure interesting local relationships (Brunsdon, Fotheringham, & Charlton, 1996)
- Purpose (research question, goals, aims, etc.)

Related Work

- Science (soil science, disease mapping)
- Public sector (social trends)
- Business? -> lack of global data

- Structure

2. Background

2.1. Simple Linear Regression

A popular technique for both explanatory and predictive data analysis is simple linear regression. Unlike “black-box” prediction tools such as Artificial Neural Networks (ANN), simple linear regression has the advantage of quantifying the individual effects of each predictor variable on the response. This quality is important for understanding why a particular outcome is predicted, as well as using regression as an exploratory tool. This general model is formalised by the following equation:

$$\boldsymbol{y_i = a_0 + \sum_{k=1}^{m} a_k x_{ik} + \varepsilon_i} \tag{1}$$

where:

y_i is the ith observation of the dependent variable y,

a_0 is the intercept,
m is the number of predictor variables,
a_k is slope/effect of the independent variable x_k,
and ε_i is the error term where ε is independent and normally distributed with mean = 0.
(Dobson & Barnett, 2008, p. 89)

For example, a linear regression model that predicts salary from years of experience and certification by professional body would include the slope of each, which can be interpreted as a predictor variable’s effect on salary with respect to the other variables in the model. So if the model was: $salary = 33 + 3(experience) + 10(certification) + \varepsilon$, then the slope/effect of experience is +3, meaning an additional year of experience at the same certification level would yield an additional 3 units of salary, plus or minus some random error, ε (see Figure 1).

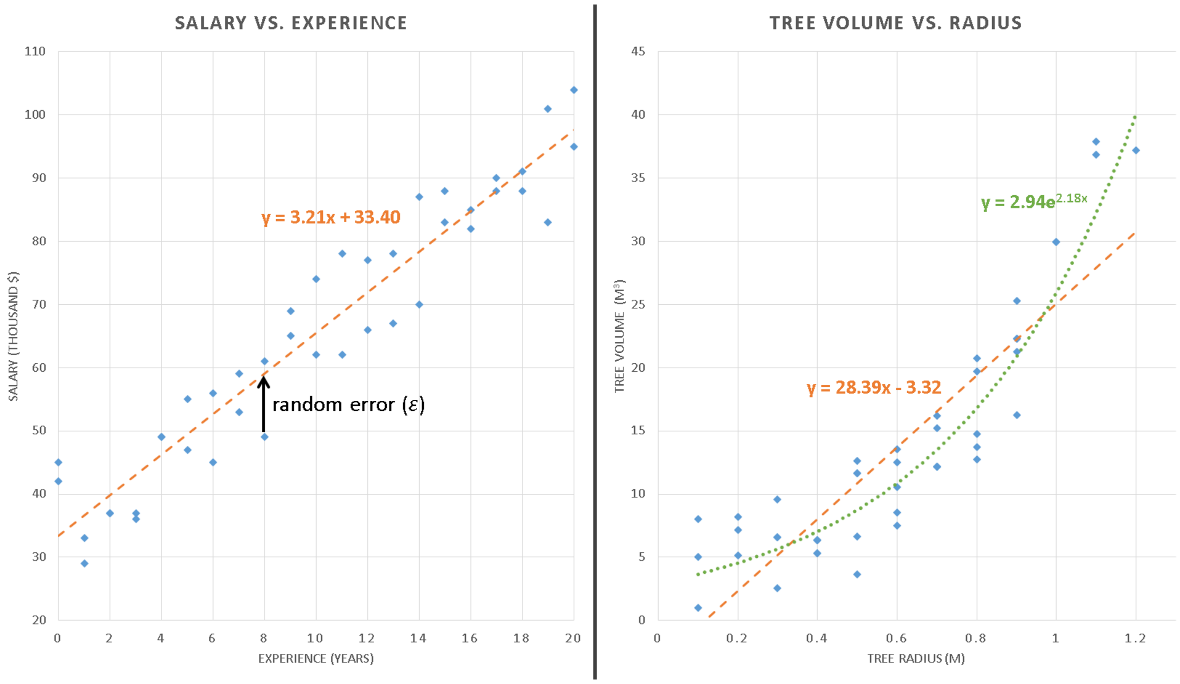

Figure 1: examples of proper (left plot) and improper (right plot) linear models using hypothetical data. Observations are depicted by blue diamonds while expected values lie along the dashed/dotted lines. The relationship between tree volume and radius (right) is clearly exponential, so the linear model (orange line) tends to underestimate volume at radii extremes and overestimate volume near average radii. The exponential model (green line) is clearly a better model. As shown in the left-hand plot, error values should be random and normally distributed about the expected value.

A key assumption in linear regression is a linear relationship between the response and each predictor. If there is a non-linear relationship, a linear model can still be fit, but the error terms (ε_is) will no longer be normally distributed, causing instability in the model (see Figure 1).

It also is important to observe that simple linear regression produces a *global* model (Brunsdon et al., 1996). Any local patterns or relationships among the data become generalized by a single line described only by an intercept and slopes for each variable. This can simplify the interpolation process where an unknown response must be predicted.

2.2. Spatial non-stationarity

The simplicity of such a global model is useful when studying truly global trends, or when the details of local phenomenon are not of interest. However, spatial data are inherently influenced by their locality. Tobler asserts that “everything is related to everything else, but near things are more related than distant things” (1970). Therefore, it makes sense to extend simple linear regression to identify interesting local relationships and how those relationships change over geographic space. These variations are referred to as spatial non-stationarity

(Brunsdon et al., 1996). To extend the global model defined by equation $\boldsymbol{y_i = a_0 + \sum_{k=1}^{m} a_k x_{ik} + \varepsilon_i}$

(1), a different model must be fit for each location of interest, formalised by the following equation:

$$y_i = a_{i0} + \sum_{k=1}^{m} a_{ik} x_{ik} + \varepsilon_i \tag{2}$$

where a_{ik} is slope/effect of the independent variable x_k *at the local observation*, i (Brunsdon et al., 1996).

This approach is referred to as Geographic Weighted Regression (GWR). However, GWR introduces additional complexity over simple regression – specifically regarding how to calculate the individual a_{ik} terms. Like simple linear regression, GWR generally uses a least squares method to find a_k terms for equation $\boldsymbol{y_i = a_0 + \sum_{k=1}^{m} a_k x_{ik} + \varepsilon_i}$

(1) that minimize S in the following equation:

$$S = \sum_{i=1}^{n} [y_i - \mu_i(a_k)]^2 \tag{3}$$

where:
y_i is the actual response at point i,
and μ_i is the expected (mean) response at point i (Dobson & Barnett, 2008, p. 14).

The minimization of error, S simultaneously for all points, i makes simple linear regression a global model. To model spatial non-stationarity then, a different model must be fit for each location, with individual errors S_i, minimized only among each i's "local" points, j. Therefore, the concept of what is considered "local" must be made explicit. In other words, defining a locality implies devising a spatial weighting function that excludes distant, non-local points by either specifying a distance buffer, or effectively reducing their weights to zero via a decay function. Brunsdon et al. recommends the following "bisquare function" which provides a balance between the lower computational cost offered by the former and the parameter surface smoothness offered by the latter:

$$w_{ij} = \left[1 - \frac{d_{ij}^2}{d^2}\right]^2 \; if\ d_{ij} < d;\ w_{ij} = 0\ otherwise \tag{4}$$

where:
i is a point where model parameters are estimated,
j is an observed point,
d_{ij} is the distance between points i and j,
and d is the spatial neighborhood distance (1996).

With this method, only points within the distance buffer, d will be considered, saving computation cost. Additionally, distant points, j near the buffer boundary will have relatively low influence on w_{ij} which will prevent the resulting parameter surface from having abrupt changes where those points are included or not (see Figure 2). Such a decay function also better respects Tobler's (1970) law by giving higher weight to closer points, j.

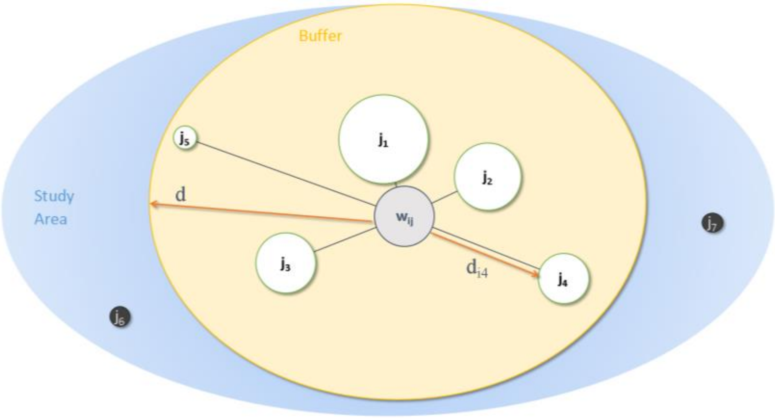

Figure 2: The spatial neighbourhood of an unmeasured parameter, w_{ik}, defined by a buffer with distance, d from w_{ik}. Nearby, measured locations (j_{1-5}) are used to estimate the value at w_{ik}. The sizes of j_{1-5} illustrate their relative weight in determining w_{ik}. Closer points are assumed to be more closely related, so they are weighted heavier. Points 6 and 7 are outside the buffer, so they are not used to calculate w_{ik}. Adapted from Mitchell (1999).

However, equation $w_{ij} = \left[1 - \frac{d_{ij}^2}{d^2}\right]^2 \; if \; d_{ij} < d; \; w_{ij} = 0 \; otherwise$ (4) assumes a fixed buffer distance, d and decay function, $[1 - \mathrm{d_{ij}^2}/\mathrm{d}^2]$. This may not be a valid assumption if the concept of "local" also varies over geographic space. Like simple linear regression, fixing the weighting function for all data may over-generalise the data and obscure local patterns. Brunsdon et al. concede that such a general weighting function may not always be appropriate, especially in economic applications where market areas vary widely (1996). For example, if a study seeks to compare customer behaviour across a large metropolitan region, market areas will likely be larger in more rural suburbs where customers are accustomed to traveling further for various services.

A possible alternative to equation $w_{ij} = \left[1 - \frac{d_{ij}^2}{d^2}\right]^2 \; if \; d_{ij} < d; \; w_{ij} = 0 \; otherwise$ (4) is a modified inverse-distance weighting (IDW) scheme. IDW is a generalized form of equation $w_{ij} = \left[1 - \frac{d_{ij}^2}{d^2}\right]^2 \; if \; d_{ij} < d; \; w_{ij} = 0 \; otherwise$ (4):

$$w_{ij} = \frac{1}{d_{ij}^{\alpha}} \tag{5}$$

where:

α is the power (distance-decay parameter),
and $\sum_i^n w_i = 1$ (Lu & Wong, 2008).

This

(4), so all points would be considered for every w_{ij} . This is computationally expensive, especially considering most points not in the vicinity of w_{ij} will be near zero. Therefore, a maximum distance buffer can be added (Lu & Wong, 2008), as in equation

$$wij = \left[1 - \frac{d_{ij}^2}{d^2}\right]^2 \; if\ d_{ij} < d;\ w_{ij} = 0\ otherwise$$

(4).

Lu and Wong (2008) extend IDW into an "Adaptive" IDW (AIDW) technique to improve interpolation over study areas in which the spatial pattern of observations vary widely, such as with the metropolitan customer study example described above. In this method, α is calculated for each unmeasured location as a function of a statistic R, defined as:

$$R = \frac{r_{obs}}{r_{exp}} \tag{6}$$

where:
r_{obs} is the average nearest-neighbor distance for w_i's 5 nearest observed points, j,
and r_{exp} is the expected (mean) nearest-neighbor distance for a random point pattern (Lu & Wong, 2008).

This statistic provides a simple means for measuring relative clustering or dispersion by comparing the observed clustering, r_{obs} to what would be expected at random, r_{exp}. The r_{exp} term only must be calculated once for the study area with:

$$r_{exp} = \frac{1}{2(n/A)^{0.5}} \tag{7}$$

where A is the area of the study region (Lu & Wong, 2008).

Finally, the R_is are mapped to values of α. Areas with high dispersion (low R) should have lower α values to give higher weight to nearby points, and vice versa; more clustered areas should not be overly influenced by very-near points (Lu & Wong, 2008). The range of appropriate α values are dependent upon the range of distance-decay values in the study area. For flexibility, Lu and Wong (2008) recommend a triangular membership function as proposed by Kantardzic (2011, p. 418) for this mapping. Equation $w_{ij} = \frac{1}{d_{ij}^{\alpha}}$

(5) can then be modified as:

$$w_{ij} = \frac{1}{d_{ij}^{\mathrm{m(R}_i)}}$$

(8)

where $m(R_{ij})$ is the mapping function from R to α.

Equation $w_{ij} = \frac{1}{d_{ij}^{\mathrm{m}(\mathrm{R}_i)}}$ (8) can now be applied to generate adaptive, neighborhood-weighted estimates of a_{ik} throughout the study area. Typically, a least-squares approach using equation $S = \sum_{i=1}^{n}[y_i - \mu_i(a_k)]^2$

(3) will be used to fit values for a_{ik} that minimize the $S_{\boldsymbol{i}}$s with respect to the neighbourhood's observations, js. The result is a surface for each coefficient (a_{ik}) in the model representing that variable's change in effect over geographic space.

3. Design

Second

3.1. Data

- Benefits (lots of data)
- Open, comparable across businesses
- Possibly biased (businesses review themselves)
- Highly subjective, noisy
- Large number of attributes

3.2. Census Data

- Source
- Factors (economic, social, racial)

3.3. Assumptions

- Smoothness of estimated coefficients (Brunsdon et al., 1996)
- Neighbourhood for estimating coefficients
- Weights of neighbours varies with each I (Brunsdon et al., 1996)
- Concept of spatial relationship (inverse square weighting)

3.4. Software tools

- Docker
- Python (pandas, geopandas, etc.)
- ArcGIS

3.5. Algorithms

- Model selection / optimisation
- Weighting function
- Statistics

3.6. Outputs

- Global model

- GWR model
- Coefficient surfaces
- Standard error surfaces (where does model perform well?)

3.7. Statistical tests

- Is GWR model significantly better than global model (Brunsdon et al., 1996)?
- Is the variation of coefficients across space significant (Brunsdon et al., 1996)?

4. Implementation

Third

4.1. Data Cleaning

4.2. Variable normalization and selection

- Linear transformations
- Sentiment analysis

4.3. Global model

- Iterative model selection

4.4. Geographic Weighted Regression model

- Iterative model selection

5. Evaluation

Fourth

5.1. Visualisation

- Smoothness
- Local patterns
- Sensitive to “edge effects” (Brunsdon et al., 1996)

5.2. Statistical significance

- Is GWR model significantly better than global model (Brunsdon et al., 1996)?
- Is the variation of coefficients across space significant (Brunsdon et al., 1996)?

5.3. Performance

- Overall: standard error, multicollinearity, etc.
- Over study area
- Processing time

6. Conclusion

Fifth

6.1. Results

- Implications (for decision makers, businesses, etc.)
- New insights

- Factors weren’t important, why? (e.g. census data)

6.2. Future work

- New questions stemming from analysis
- Enhancing models
- Better neighbourhood estimators (alternative to IDW, e.g. drive time vs. circular buffers)

References

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis, 28*(4), 281-298.

Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models*: CRC press.

Kantardzic, M. (2011). *Data mining : concepts, models, methods, and algorithms* (Second Edition.. ed.): Piscataway, New Jersey : IEEE Press ; Hoboken, NJ : Wiley. 2011 ©2011.

Lu, G. Y., & Wong, D. W. (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Computers & Geosciences, 34*(9), 1044-1055.

Mitchell, A. (1999). *The ESRI Guide to GIS Analysis: Geographic patterns & relationships* (Vol. 1): ESRI, Inc.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography, 46*(sup1), 234-240.

Appendix