The Impact of Varying Semantics in Spatial Statistics

Comber AJ[*1], Atkinson PM[†2] and Harris P[‡3]

[1] School of Geography, University of Leeds
[2] Faculty of Science and Technology, Lancaster University
[3] Sustainable Agricultural Sciences, Rothamsted Research

January 12, 2018

Summary

This paper extends consideration of semantics to identify the sources of semantic variation in modelling, using a case study on soil conductivity in the Loess Plateau, China through spatial regression as the model. Given a simple case of $y = f(x)$, noting that y will only ever be a predicted, y^*, the paper examines the impacts on semantics of *Measurement of x* (the **'x' issue**), the *Choice of x* (the **'m' issue)** and the *Support of x* (the **'v' issue**). These affect the semantics of y^* through model inputs, specification and granularity, and although some of these factors are known, they have not been considered in this context together before.

KEYWORDS: sample measurement; model specification, sample support, spatial regression.

1. Introduction

The three key areas that Geography contributes to the wider scientific community are representation, scale and uncertainty. How we represent a process is critical. And whilst consideration of spatial data semantics is well developed – the way that real-world features are conceptualized and represented in our databases – very little work has examined the semantics of the way we analysis processes in our statistical models – what we *do* with our spatial data. Specifically, the semantics associated with how we model *spatial processes* is an under-researched area. This paper explores the semantics associated with constructing statistical models of spatial processes and identifies the sources of semantic variation in modelling.

Taking the simplest case of y as a function of x:

$$y = f(x) \tag{1}$$

and noting that y will only ever be an approximation, y^*, this paper examines the semantic impacts on statistical models of y arising from decisions over:

- *Measurement of x*: the **'x' issue** in which different measures of 'x' influence the model of y^*;
- *Choice of x*: the **'m' issue** of what and how many predictors to include in the specification of the model itself which has a semantic effect on y^*;
- *Support of x*: the **'v' issue** in which support effects are induced through the choice of measurements scales and granularity.
-

The first two (the **'x' and 'm' issues**) are reasonably well-known within the statistics and modelling communities and many tests exist to determine which measures of x to use and which predictors to select. However, the *Support* (the **'v' issue**) is less frequently considered and yet has profound effects on the resulting model semantics (i.e., how y^* is specified). This includes the geographic space on which predictors are defined with different resulting integrals of supports for y^*. Support effects are

[*] a.comber@leeds.ac.uk
[†] pma@lancaster.ac.uk
[‡] paul.harris@rothamsted.ac.uk

induced through the choice of model, particularly for spatial models. The impacts of these issues are illustrated by modelling soil conductivity in the Loess Plateau, China.

2. Case study, data and models

For this study, soil conductivity (*y*) was hypothesised as being described by bulk density, field capacity, soil saturated water content, elevation, slope, aspect, stream flow, topographic position, ruggedness and land surface roughness. Other predictor variables are frequently included as described in Liu et al (2008). Field data were collected at 224 locations in the Loess Plateau, China (Figure 1). The data are described in Zhao et al (2016). For each sample, several soil properties were measured (soil conductivity, soil bulk density and soil saturated water content) at different depths of 0 to 10 cm, 10 to 20 cm and 20 to 40 cm. The different measurements of these *x* variables (taken as the **'x' issue**) have different values at the same sites and different distributions (Figure 2).

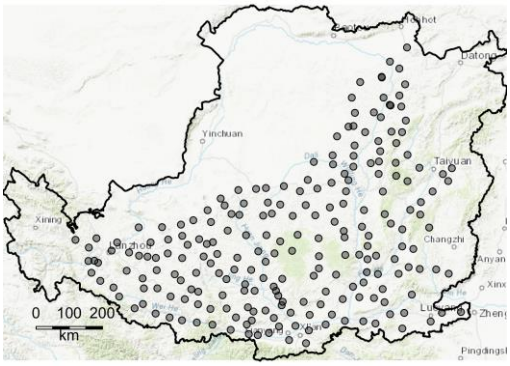

Figure 1 The field sites and the Loess Plateau study area.

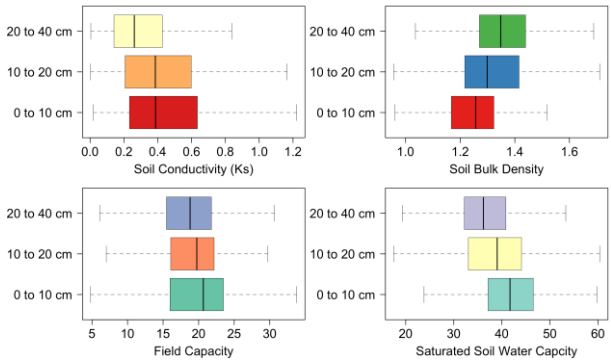

Figure 2 Variation in the different measurements of *x* from field data.

In addition, a 30 m DEM was resampled to 90 m and 180 m. Terrain variables were generated (aspect, stream flow direction, slope, Topographic Position Index (TPI), Terrain Ruggedness Index (TRI) and surface roughness). Spatial intersection was used to extract the terrain variables for each of the sample sites. These terrain predictor variables have different distributions at different resolutions (Figure 3) indicating variation in the support of *x* (the **'v' issue**) from data collected at different resolutions, scales and granularity.

The final source of variation in *y** is the predictors that are used as inputs to construct the model. This is the **'m' issue** in which the specification of the model itself has a semantic effect on *y**. In the general case, this includes what and how many predictors (*x)* to include in the model and how to deal with the error term. The simplest model is a global regression as follows:

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \varepsilon_i \qquad (2)$$

where, for observations indexed by $i = 1,\ldots,n$, y_i is the response variable, x_{ij} is the value of the j^{th} predictor variable, m is the number of predictor variables, β_0 is the intercept term, β_j is the regression coefficient for the j^{th} predictor variable and ϵ_i is the random error term. As a spatial regression model, geographically weighted regression (GWR) is chosen, which is similar to the global case but calculates a series of local linear regressions with locations associated with the coefficient terms:

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{m} \beta_j(u_i, v_i) x_{ij} + \varepsilon_i \qquad (3)$$

where (u_i, v_i) is the spatial location of the i^{th} observation and $\beta_j(u_i, v_i)$ is a realization of the continuous function $\beta_j(u, v)$ at point i. The geographical weighting results in data nearer to the kernel centre making a greater contribution to the estimation of local regression coefficients at each local regression calibration point k. For this study, the weights were generated using a bi-square kernel, which for the bandwidth parameter r_k is defined by:

$$w_{ik} = \left(1 - \left(d_{ik}/r_k\right)^2\right)^2 \text{ if } d_{ik} \le r_k \quad w_{ik} = 0 \quad \text{otherwise} \tag{4}$$

where the bandwidth can be specified as a fixed (constant) distance value, or in an adaptive, varying distance way, where the number of nearest neighbours is fixed (constant). In this case, adaptive bandwidths were chosen.

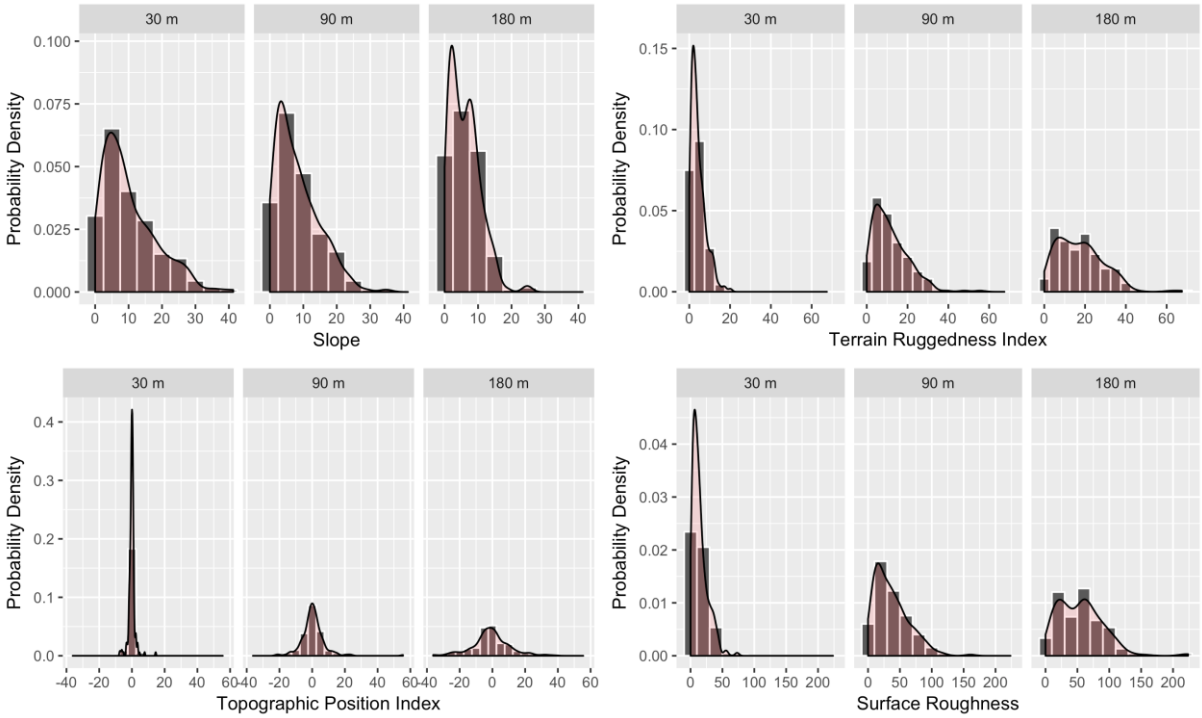

Figure 1 Variation in the support of *x* from the terrain analysis at different resolutions.

3. Results

A general model of soil conductivity was hypothesised as being described by the following predictors: bulk density, field capacity, soil saturated water content, elevation, slope, aspect, stream flow, topographic position, ruggedness and land surface roughness. Nine global regression models were fitted to model soil conductivity at three different depths, using field data predictors measured at those depths (the **'x' issue**) plus the terrain predictors created from DEM data at different resolutions (the **'v' issue**). Each of these models was then subjected to a step-wise AIC model selection procedure (the **'m' issue**), resulting in parsimonious models with a reduced set of (selected) predictors (Table 1).

The 18 regressions were compared using an analysis of variance (ANOVA). The reductions in the residual sum of squares (RSS) between each of the models is shown in Table 2. There are some broad patterns. While no significant differences were found between models 1 to 9, significant differences were found between models 4 to 9 and selected models, for example:

- *40 cm depth 90 m DEM* (model 8) and the selected *10 cm depth 30 m DEM* (model 1ms), with reductions in the RSS of 18.93
- *10 cm depth 90 m DEM* (model 2) and the selected *40 cm depth 30 m DEM* (model 7ms) with reductions in the RSS of 19.00
- *40 cm depth 180 m DEM* (model 9) and the selected *20 cm depth 90 m DEM* (model 5ms) with reductions in the RSS of 3.96

A GWR analysis of the 6 highlighted global models above, generated alternative predicted measures of *y*, *y** which are compared with observed *y* in Figure 4. Here the impacts of the different model semantics begin to be evident: they are each generating predictions of *y** that have different spatial, measurement and model selection characteristics.

Table 1 Selected predictors for each model.

Model ID	Model
1	10 cm depth 30 m DEM
2	10 cm depth 90 m DEM
3	10 cm depth 180 m DEM
4	20 cm depth 30 m DEM
5	20 cm depth 90 m DEM
6	20 cm depth 180 m DEM
7	40 cm depth 30 m DEM
8	40 cm depth 90 m DEM
9	40 cm depth 180 m DEM

Model ID	Selected predictors
1ms	SSWC.0.10cm.
2ms	SSWC.0.10cm.
3ms	SSWC.0.10cm. roughness slope
4ms	SSWC.10.20cm. slope elev.
5ms	SSWC.10.20cm. roughness elev.
6ms	SSWC.10.20cm. elev.
7ms	BD.20.40cm. roughness slope
8ms	BD.20.40cm. FC.20.40cm. elev.
9ms	BD.20.40cm. FC.20.40cm. elev.

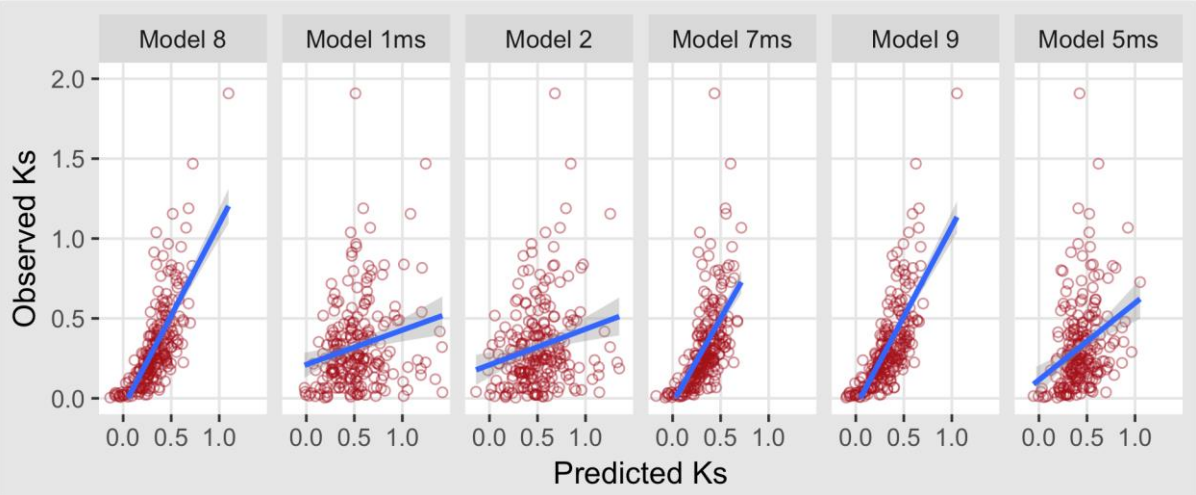

Figure 4 Observed (*y-axis*) against Predicted (*x-axis*) soil conductivity from 6 selected models.

The impacts of model semantic variation are further exemplified when the predictions from GWR are mapped over a 10km grid covering the study area as in Figure 5. Here, the ratios of predicted to observed soil conductivity are mapped to show how and where disagreement varies spatially, with the size of the adaptive bandwidth giving an indication of spatial heterogeneity and the degree with which the semantics of modelled soil conductivity vary *within* the models as well as between them.

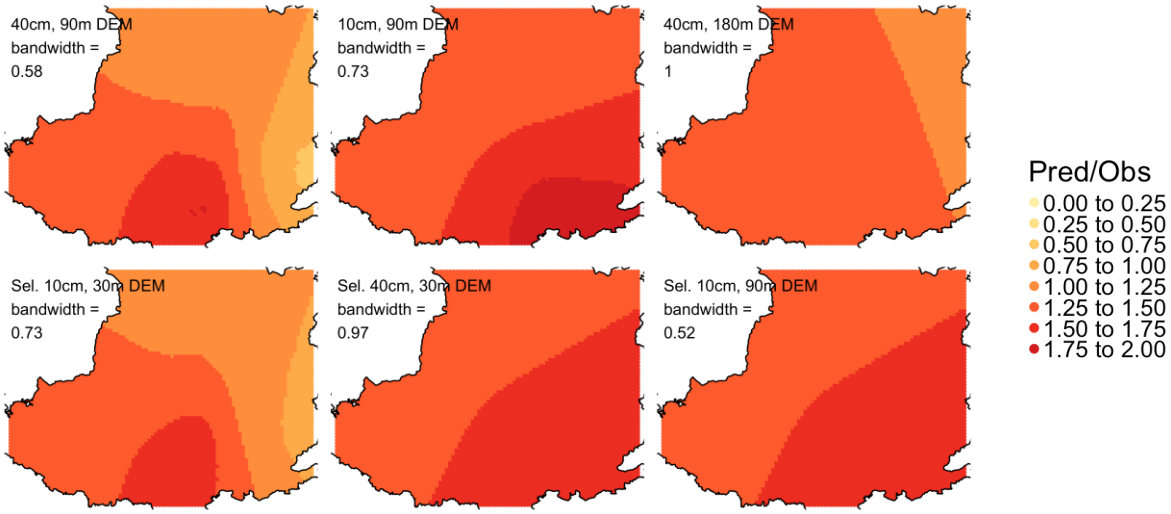

Figure 5 GWR surfaces of the ratios of predicted to observed soil conductivity for 6 selected models.

4. Discussion

Statistical models are fitted to describe, understand and predict various processes. In Geography our models may be spatial, and spatial data are frequently used as model inputs. Model semantics are defined by **x**, **v** and **m** issues. The impact of **x** and **m** are well known within statistical communities. However, their interaction with the **v** issue is less well understood and frankly under considered. Consider the GWR results in Figures 4 and 5: these show how well predicted matches observed soil conductivity. Each of the models has different choices of *x*, (**x-issue**), the 'ms' models have different selections of *x* (**m-issue**) and different spatial supports for *x* (**v-issue**). Each of these fundamentally affects the meaning and semantics of predicted / modelled *y* in global (fixed coefficient) models. The potential for such semantic differences is further enhanced when spatial regression models, such as GWR, are used as the coefficient estimates themselves, also vary across space. The result is that the *local* regression model predicting soil conductivity is different in different places, adding an additional layer to the semantic variation inherent in global regressions.

5. Acknowledgements

This research was supported by the Natural Environment Research Council Newton Fund grant (NE/N007433/1) and a UK Biotechnology and Biological Sciences Research Council grant (BB/J004308/1). Analyses and mapping were undertaken in R 3.3.3 - open source statistical software.

6. Biography

Lex Comber holds a Chair in Spatial Data Analytics and works at the University of Leeds.
Pete Atkinson is Dean of Faculty of Science and Technology, Lancaster University.
Paul Harris enjoys stats and works for Rothamsted Research as an Eco-Informatics Scientist.

References

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

Liu, J., Pattey, E., Nolin, M.C., Miller, J.R. and Ka, O., 2008. Mapping within-field soil drainage using remote sensing, DEM and apparent soil electrical conductivity. Geoderma, 143(3):261-272.

Zhao, C., Shao, M.A., Jia, X., Nasir, M. and Zhang, C., 2016. Using pedotransfer functions to estimate soil hydraulic conductivity in the Loess Plateau of China. *Catena*, 143:1-6.

Table 2 Reductions in the residual sum of squares between each of the models (significant differences are in bold)

Model	1	2	3	4	5	6	7	8	9	1MS	2MS	3MS	4MS	5MS	6MS	7MS	8MS	9MS
1	0.00																	
2	0.40	0.00																
3	-0.40	-0.80	0.00															
4	-14.44	-14.85	-14.05	0.00														
5	-14.28	-14.69	-13.89	0.16	0.00													
6	-14.24	-14.65	-13.85	0.20	0.04	0.00												
7	-18.56	-18.96	-18.16	-4.11	-4.27	-4.31	0.00											
8	-18.12	-18.52	-17.72	-3.67	-3.83	-3.87	0.44	0.00										
9	-18.21	-18.62	-17.81	-3.77	-3.93	-3.97	0.34	-0.09	0.00									
1MS	0.81	0.41	1.21	**15.26**	**15.10**	**15.06**	**19.37**	**18.93**	**19.03**	0.00								
2MS	0.81	0.41	1.21	**15.26**	**15.10**	**15.06**	**19.37**	**18.93**	**19.03**	0.00	0.00							
3MS	0.04	-0.37	0.44	**14.48**	**14.32**	**14.28**	**18.59**	**18.16**	**18.25**	-0.78	-0.78	0.00						
4MS	-14.26	-14.66	-13.86	0.19	0.03	-0.01	**4.30**	**3.86**	**3.96**	**-15.07**	**-15.07**	-14.29	0.00					
5MS	-14.25	-14.65	-13.85	0.19	0.03	-0.01	**4.31**	**3.87**	**3.96**	**-15.06**	**-15.06**	-14.29	0.01	0.00				
6MS	-14.09	-14.49	-13.69	0.35	0.20	0.15	**4.47**	**4.03**	**4.12**	**-14.90**	**-14.90**	-14.13	0.17	0.16	0.00			
7MS	-18.18	-18.59	-17.78	-3.74	-3.90	-3.94	0.37	-0.07	0.03	**-19.00**	**-19.00**	-18.22	-3.93	-3.93	-4.09	0.00		
8MS	-17.91	-18.31	-17.51	-3.47	-3.63	-3.67	0.65	0.21	0.30	**-18.72**	**-18.72**	-17.95	-3.65	-3.66	-3.82	0.27	0.00	
9MS	-17.91	-18.31	-17.51	-3.47	-3.63	-3.67	0.65	0.21	0.30	**-18.72**	**-18.72**	-17.95	-3.66	-3.66	-3.82	0.27	0.00	0.00