Article:
Comber, A (2019) The forgotten semantics of regression modelling in Geography. Geographical Analysis. ISSN: 0016-7363

https://doi.org/10.1111/gean.12199

The forgotten semantics of regression modelling in Geography

Comber AJ[1], Harris P[2], Lü Y[3], Wu L[2] and Atkinson PM[4, 3, 5, 6]

[1] School of Geography, University of Leeds, UK, email: a.comber@leeds.ac.uk

[2] Sustainable Agriculture Sciences, Rothamsted Research, UK, email: paul.harris@rothamsted.ac.uk; lianhai.wu@rothamsted.ac.uk

[3] Chinese Academy of Sciences, Beijing, 100085, China, email: lyh@rcees.ac.cn

[4] Faculty of Science and Technology, Lancaster University, UK, email: pma@lancaster.ac.uk

[5] School of Geography and Environmental Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK

[6] School of Natural and Built Environment, Queen's University, Belfast, Belfast BT7 1NN, UK

Abstract

This paper is concerned with the semantics associated with the statistical analysis of spatial data. It takes the simplest case of the prediction of variable *y* as a function of covariate(s) *x*, in which predicted *y* is always an approximation of *y* and only ever a function of *x*, thus, inheriting many of the spatial characteristics of *x*, and illustrates several core issues using 'synthetic' remote sensing and 'real' soils case studies. The outputs of regression models and, therefore, the meaning of predicted *y*, are shown to vary due to 1) choices about data: the specification of *x* (which covariates to include), the support of *x* (measurement scales, granularity), the measurement of *x* and the error of *x*, and 2) choices about the model including its functional form and the method of model identification. Some of these issues are more widely recognised than others. Thus, the study provides definition to the multiple ways in which regression prediction and inference are affected by data and model choices. The paper invites researchers to pause and consider the semantic meaning of predicted *y*, which is often nothing more than a scaled version of covariate(s) *x*, and argues that it is naïve to ignore this.

Keywords

Spatial data capture; spatial statistics; non-stationary; autocorrelation; scale; model selection

1. Introduction

There is a long and varied literature on the semantics associated with geographical information. Primarily this is concerned with how spatial phenomena are conceptualised,

represented and encoded in spatial data (e.g. Harvey, 2000; Smith, 2001; Smith and Varzi, 2000; Smith and Mark, 2003; Pires, 2005; Mark and Turk, 2003; Comber, 2005; Turk et al., 2011; Derungs et al., 2013; Robbins, 2001). Much less attention has been paid to the semantics associated with the analysis of spatial data, that is, what we *do* with spatial data. In this paper, we describe how the specification of spatial models, and choices made therein, have an impact on the outcomes of a spatial analysis. For example, taking the simplest case of the prediction of y as a function of x with error ε:

$$y = f(x) + \varepsilon \qquad (1)$$
$$y^* = f(x)$$

and noting that the predicted y, i.e., y^*, will only ever approximate y, it is evident that y^* in Equation 1 is always, and can only ever be, a function of x. Thus, only $f(x)$ is ever found and never y. This statement is axiomatic but often overlooked. It can lead to implicit or explicit claims about regression-based spatial outputs that are indefensible. For example, maps of disease incidence have been produced that are in effect simple transforms of covariates such as vegetation greenness (e.g., Beck et al., 2000; Anyamba et al., 2001; Hay et al., 2005; Lui et al., 2015). In these, the disease map is actually a map of vegetation greenness scaled by additive and multiplicative factors. The map inherits all the spatial characteristics of vegetation greenness and only those of disease incidence that relate indirectly through greenness. Thus, the key points that this paper makes are that y^* can only ever be a function of x; that only $f(x)$ is ever found and never y; and that models built on these paradigms, including those using spatial data, may result in naïve interpretations of the results and rather surprising consequences in some extreme cases.

The semantic and characteristic variation in y^*, as a function of how y is predicted, will arise due to core decisions about the data and the model. Decisions relating to the data include their specification, support, measurement and errors:

- The *specification of x*: the ***x* issue**, on which covariates to include in the model.
- The *support of x*: the ***v* issue** in which support effects are induced through the choice of measurement scales and granularity of x.
- The *measurement of x*: the ***m* issue**, that is, how the desired x is measured.
- The *error of x*: the ***e* issue** in which the accuracy with which we measure x, affects y^*.

Decisions relating to the model include its functional form, the method of model identification and the specification of *y* itself (a combined data and model issue).

- The *functional form of the model*: the ***f* issue,** in which the specification of the model itself has a semantic effect on *y**.
- The *method of model identification*: the ***i* issue**, the statistical method of estimating the model parameters.
- The *specification of y*: the ***y*** issue. For any model incorporating spatial effects, semantic and characteristic variation in *y** will also arise as a function of the specification of *y*. The definition and sampling framework of *y* itself has a semantic effect on *y** at locations other than *y*.

This study illustrates the effects of data and model decisions through the ***x***, **v,** ***m*****,** ***e***, ***f*** and ***i*** **issues** and their impacts on the semantics of the resultant spatial operations. Focus is placed on the imparted effects on *y** through these issues (i.e., model prediction), and much less so the imparted effects on associated inferences (i.e., parameter uncertainty). The ***y*** **issue** is acknowledged, but not investigated, as its illustration is not straightforward and could not be captured adequately here (see Wang et al. (2012) and references therein). Although the stated issues are inherent to many statistical operations, we focus on regression, where the issues are illustrated through ordinary least squares (OLS) multiple regression, geographically weighted regression (GWR) (Brunsdon et al., 1996), and mixed models whose parameters are estimated via restricted maximum likelihood (REML) (e.g., Welham et al., 2004) or via maximum likelihood (ML). The next section describes the six issues in more detail before they are illustrated using a 'synthetic' case study of Landsat 5 remote sensing imagery (Section 3). The analysis is repeated using a 'real' case study of a soils dataset for the Loess Plateau, China. Finally, we conclude the study with key outputs and discussion points.

2. Semantic issues in detail

Regression seeks to model changes in *y*, the target variable, with changes in *x*, the covariates, where *x* is always (and naively) assumed free of measurement error. For this study, *y* is always fixed in fitting the regression, meaning that regardless of the issues described, the aim is always towards the same *y*, producing a *y** with associated residuals (*y* - *y**). As we are not presenting the **issue *y***, only *in-sample* prediction of *y* is conducted and not *out-of-sample*.

2.1 The *x* issue

The ***x* issue** as relates to the number k_x and choice c_x of potential covariates to present to the regression. Covariate selection has a limiting effect on the possible value and semantic meaning of y^*. Covariate *selection* can be a difficult task. For example, two or more covariates displaying similar predictive ability in relation to y can commonly be retained to increase the accuracy of y^*, but at the same time can compromise model interpretability due to their collinearity. (Note that presence of such covariance amongst covariates can confound the semantics of any model due to uncertainty about whether the observed response is a result of x_1 or of x_2, imposing a further level of intra-x semantics on the model.) This duality can result in an ***i* issue** of whether to fit the model using, say, some penalized regression or not (e.g., Zou and Hastie 2005). Covariate *availability*, links to the ***f* issue** (and indirectly to the ***i* issue)**, as missing covariates in a spatial study are often reflected in spatial effects (e.g., non-stationary relationships and/or autocorrelated residuals), that would not be present if the regression were fully-specified (i.e., with a full set of covariates, such that a multiple regression fitted using OLS suffices) (see Cressie and Chan 1989).

2.2 The *v* issue

The support ***v*** of x is an important concept in geostatistics and quantitative geography. The support is the space on which an observation is defined, or on which a measurement is made and is defined by parameters such as size, shape, orientation, position etc. as well as the dimension of interest (e.g., 2D, 3D, 2D plus time). The support represents the lower limit of what we can know about the real world. In contrast, the spatial extent of the sample set defines the space of interest, while the sampling scheme (e.g., random, systematic grid) defines how that space is represented. Nothing can be known *beyond* the extent of the sample set and nothing can be known *within* the support (Atkinson and Tate, 2000). Indeed, all that can be known exists in the relations *between* the observations of the specific sampling scheme. As a result, data are always a function of the real world and the sampling framework, as well as conceptual and semantic choices defining measurable properties (such as ‘biomass per unit area’). The support has an important effect on the semantic meaning of any regression prediction: as the support increases, some of the potential variation in the property of interest is lost to within the support (i.e., it is integrated out) and all that is left is the variation between supports. Because of this, the measured variance collapses as the support size increases.

Where spatial structure exists, neighbouring observations are commonly similar to each other (positive spatial correlation) or more rarely dissimilar (negative spatial correlation), entailing in both instances that decrease in variance is smaller as the support increases. This means that less variance is lost to within the support. This is also one aspect of the modifiable areal unit problem (Openshaw, 1984), where, as the areal size changes (up- or down-scaling), the correlation between a given pair of variables changes, conditional upon the scales of spatial variation that exist in the same variables. Gotway and Young (2002), Zhang et al. (2014) and Murakami and Tsutsumi (2015) describe the effects of varying support of the observation units.

In terms of the ***v* issue**, y^* and $(y - y^*)$ are a function, either directly or indirectly, of the sampling framework of x, specifically, the size, geometry, and orientation of the support over which the variables were measured. If the supports of the covariates x change, then the predictions y^* will change and, thus, the semantics of the model or its predicted state will change. Most obviously, the variance of y^* will change.

2.3 The *m* issue

The ***m* issue** can manifest itself in several ways. Measurements of a property will vary according to different options for how they are measured analytically. If these variables are then used separately as covariates to predict y, then clearly y^* will vary according to which form of the covariate x is specified. Thus, the specific impact of any covariate x is dependent on its method of measurement, and each individual covariate may be affected in this manner in different ways.

An interesting ***m* issue** consideration arises in remote sensing. The pixel values in a remotely sensed image are generally accepted to be integrals (e.g., of radiance or brightness) over the measurement support, which is approximated as a pixel. The measured values are integrals across different variables because the angle of view varies across the support. This may matter less for certain satellite-based observations, but it may be of greater consequence for airborne, drone or ground-based measurements. For example, when measuring the radiance of a crop using a ground-based instrument, a single measurement is likely to include the tops of the canopy at nadir, but also the sides of the canopy at the edge of the support. Thus, a

single value is realised as the integral of fundamentally different variables. This semantic mixing in x will impact on y^*.

2.4 The *e* issue

The ***e* issue** is concerned with measurement error and how the accuracy with which we measure x, affects y^*. The x's can never be known perfectly and while they *might* be error free, we can never know that. Thus, the error e_x in x will propagate through the model transform to y^*. Here e_x may include systematic and random contributions of given magnitudes, will have its own distributional form, and may have a spatio-temporal character (e.g., autocovariance). It may also be related directly to x (i.e., heteroscedasticity). All of these properties will impact on y^*, and regression models to account for them exist (e.g., Christensen, 2011). Most fundamentally, the larger the error e_x relative to the signal in x (i.e., the smaller the signal-to-noise ratio in x), the more degraded is the semantic information content of y^*. Observe that the error e_x is different to the error or residual of y^*.

2.5 The *f* issue

The ***f* issue** is concerned with model specification or choice. For regression modelling this can include models that deal with non-linearity or heteroskedasticity, relationship non-stationarity (e.g., GWR) and autocorrelated residuals (e.g., mixed models). For each model category there exist alternative models (or modelling paradigms) with broadly the same objective. For example, weighted regression can be used to deal with non-linearity or heteroskedasticity (e.g., Carroll and Rupert 1988), an expansion model can be used to model non-stationary relationships (Cassetti, 1972), and a simultaneous autoregressive model can be used to counter autocorrelated residuals (e.g., Anselin, 1990; Cressie, 1993).

Model specification frequently includes secondary choices within the chosen model. These include the kernel function in GWR (e.g., Gaussian, bi-square, etc.) or the variogram model in a mixed model (e.g., exponential, spherical, etc.). All these choices can be directed based on model fitting, but only where the options have been explored by the investigator. This is analogous to the case for the data (the ***x* issue**) whereby the data selected by the model fitting can be chosen only from the set provided by the investigator: the chosen model can only be from the set of models investigated.

2.6 The *i* issue

Model identification concerns the statistical estimation method used for fitting a chosen model. Examples include weighted least squares (WLS) for addressing non-linearity or heteroskedasticity and partial least squares (PLS) for addressing collinearity (e.g., Frank and Friedman, 1993). The method of fitting can also determine the number $k_{\mathbf{x}}$ and set $c_{\mathbf{x}}$ of covariates to include in the model and the use of an intercept term (e.g., with stepwise, LASSO or elastic net approaches; see Zou and Hastie, 2005). The choice of adopting a Bayesian inference framework also falls mostly into this category, for example, through fitting options such as Markov chain Monte Carlo (MCMC) or Integrated Nested Laplace Approximation (INLA) methods (Rue et al., 2009). The choice of estimation method will determine the identification of the model and parameters specified and, thus, this choice will have an effect on y^*, its semantic meaning and character.

The method of identification strongly links with model specification. However, we do not attempt to illustrate more complex problems in which it can be difficult to isolate non-linearity, non-stationarity and autocorrelation effects, for a given spatial dataset (e.g., Anselin, 2010; Basile et al., 2014; Harris, 2018) or illustrate those that additionally consider scale-dependent processes (e.g., Dong et al., 2015). Instead, we choose to illustrate routine methods of identification in mixed models with REML and ML, both of which separate trend (first-order effects) from autocorrelation (second-order) effects (see also, Armstrong, 1984).

2.7 Summary

It is clear that the semantic meaning of the regression-based prediction of y^* is a function of the number k_x and choice c_x of covariates x (the ***x* issue**), the given supports (the ***v* issue**), the way that they are measured (the ***m* issue**), together with their (unknown) errors (the ***e* issue**). The fitted regression in terms of its specification (the ***f* issue**) and its identified parameters (the ***i* issue**) will also impact on y^*. These issues and, in particular, their impact on the prediction of y^* and model inference, are illustrated through two case studies with *non-spatial* multiple regression models and GWR, and mixed models as *spatial* regression models. GWR and the mixed model represent the ***f* issue,** and the ***i* issue** is represented by OLS for multiple regression, together with REML/ML for the mixed models. All study regression models are described in the Appendix to this paper.

3. Synthetic case study: Landsat 5 imagery

In the first case study, remote sensing imagery is used to illustrate the semantic impacts of regression modelling on y^* and model inference due to the ***x***, **v, *m*, *e*, *f*** and ***i*** **issues**. Cloud-free Landsat 5 imagery from 2011 was obtained for the East of England (path 201, row 023, 30th October 2011) with a spatial resolution of 30 m. Three of the Landsat 5 spectral bands were used in the regression analyses, with the Blue band (0.45-0.52 μm) as the target variable (y), and the Red and Green bands (0.63-0.69 μm and 0.52-0.60 μm respectively) as the covariates x_1 and x_2, respectively. A 50 x 50 pixel area was selected randomly and extracted for use (Figure 1).

Figure 1. The 50 x 50 pixel area of 2011 Landsat 5 imagery used in the study, showing bands 3, 2 and 1 through the Red, Green and Blue channels.

A total of eight regression analyses were undertaken to illustrate the semantic effects of the six issues, where in each case the aim was to predict the Blue band (y) using the Red band (x_1) and in three instances, the Green band (x_2) also. The eight regression analyses were as follows:

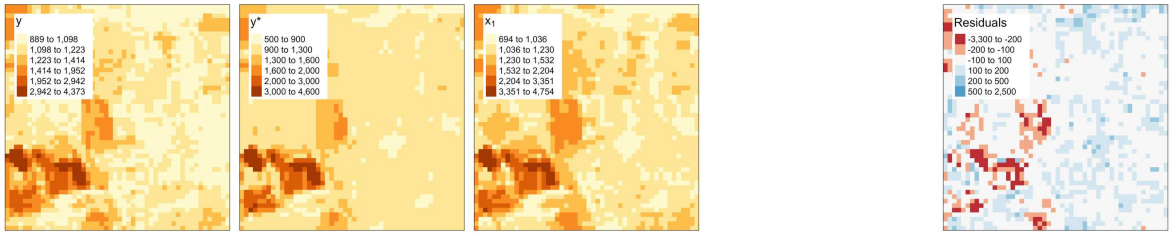
a) **Reference:** $y = f(x_1)$ with multiple regression;
b) ***x* issue:** $y = f(x_1 + x_2)$ with multiple regression;
c) ***v* issue:** $y = f(x_1)$ with multiple regression, where the Red band (x_1) is aggregated to 60 m spatial resolution using a nearest neighbour algorithm;
d) ***m* issue:** $y = f(x_1)$ with multiple regression, where the Red band (x_1) is altered slightly by taking its square root;
e) ***e* issue:** $y = f(x_1)$ with multiple regression, where a relatively small amount of random noise is added to the Red band (x_1);
f) first ***f* issue:** $y = f(x_1)$ with GWR;

g) second ***f* issue:** $y = f(x_1 + x_2)$ with a mixed model fitted by REML;

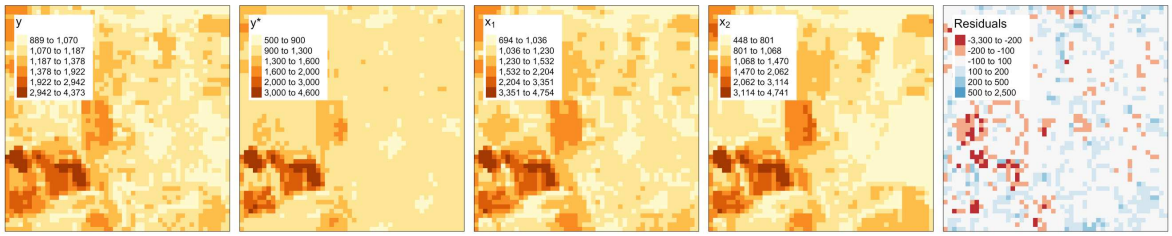
h) ***i* issue:** $y = f(x_1 + x_2)$ with a mixed model fitted by ML.

Figure 2 maps the actual y's (all the same), the different predicted y^*'s, the covariates x_1 and x_2, and the residuals (y - y^*) arising from the eight regression fits and the six semantic issues ***x***, **v**, ***m***, ***e***, ***f*** and ***i***. The distributions of y^* are also shown via boxplots in Figure 3, along with the corresponding residuals (y - y^*) in Figure 4. It is also possible to re-compare the residuals ($y - y^*$) as in Figure 5, together with y^*, with the covariate x_1 through ($x_1 - y^*$) as in Figure 6, where maps from both figures use the same legend and class breaks. Model inference for each of the eight regression models is summarised in Table 1.

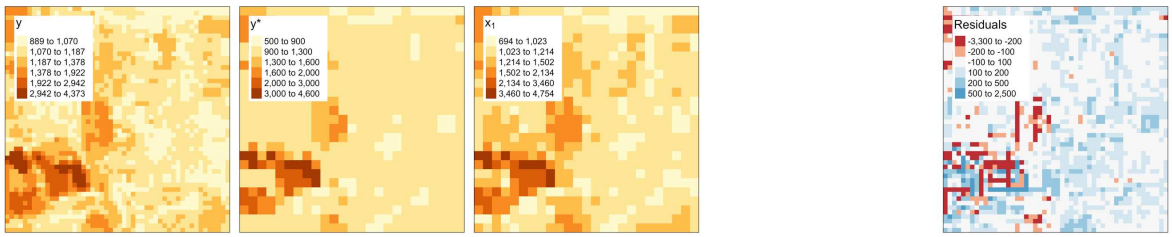
a) **Reference:** multiple regression with the Red band as the single covariate



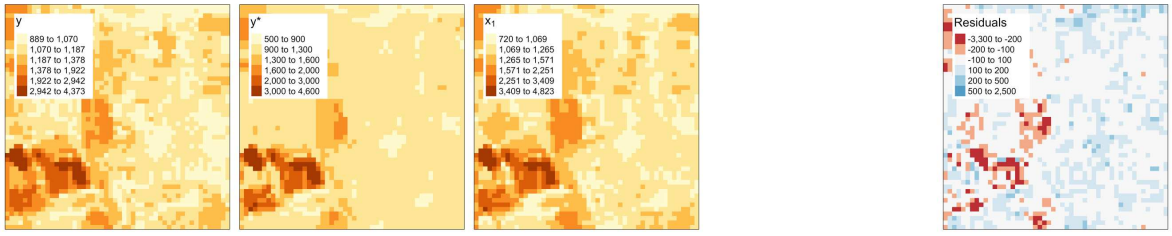
b) The ***x* issue:** multiple regression with the Red and Green bands as covariates



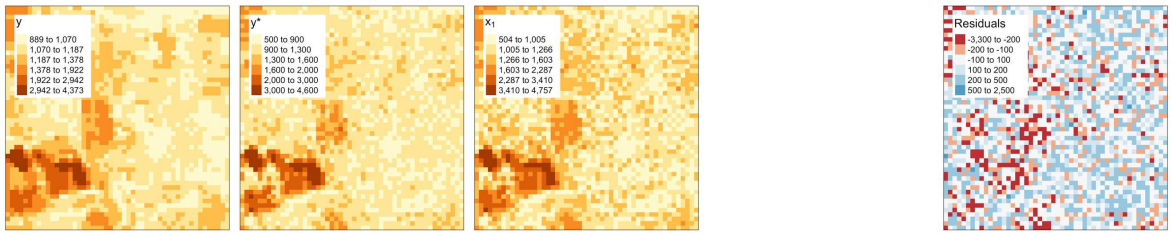
c) The ***v* issue:** multiple regression with the Red band aggregated to 60m



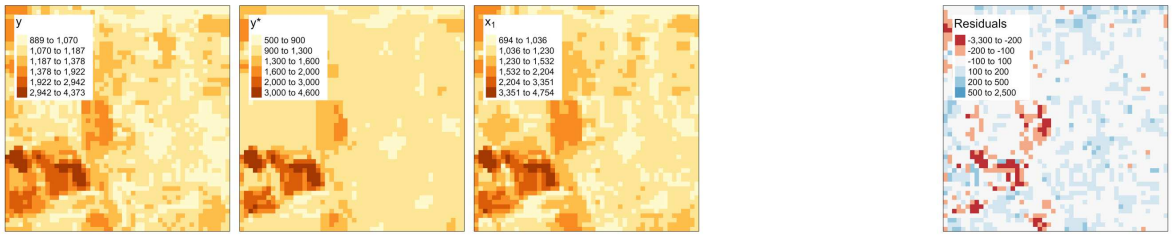
d) The ***m* issue:** multiple regression with the Red band altered, by taking its square root



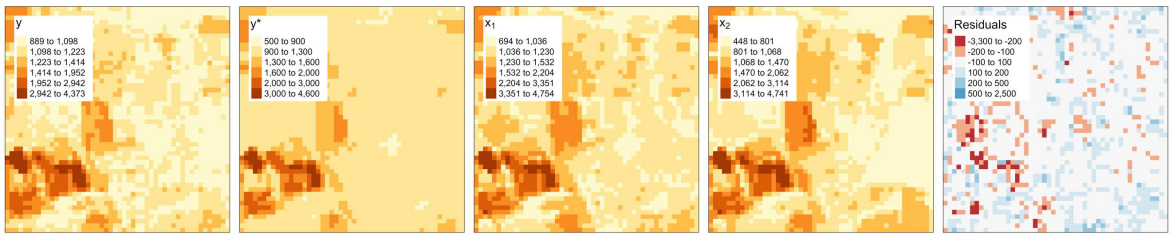
e) The ***e* issue:** multiple regression where random noise is added to the Red band



f) The first ***f* issue:** GWR with the Red band as the single covariate



g) The second ***f* issue**: mixed model REML with the Red and Green bands as covariates



h) The ***i* issue**: mixed model ML with the Red and Green bands as covariates

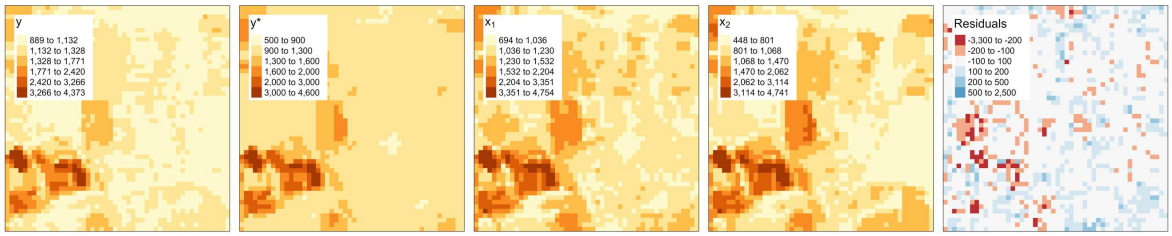


Figure 2. Maps of the original and same target variable (y), different y*, covariates (x_1 and x_2) and model residuals ($y - y$*) arising from the six issues and eight regressions (a to h).

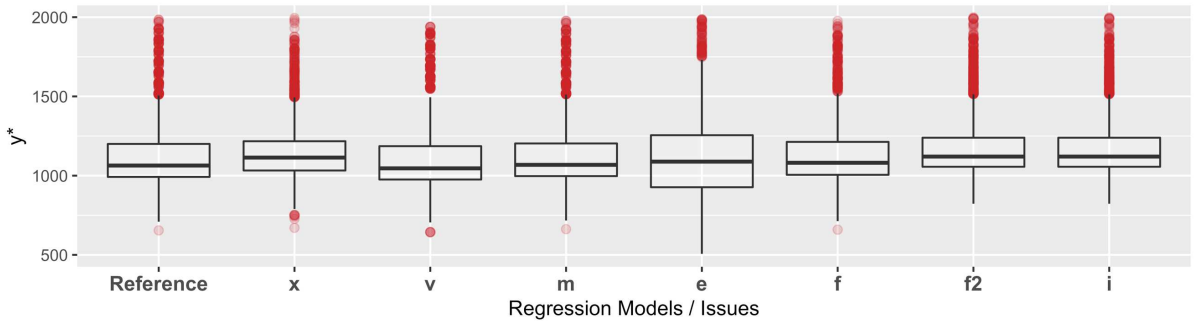


Figure 3: Boxplot distributions of y* arising from the six issues and eight regressions (a to h) (outliers given with a small transparency term).

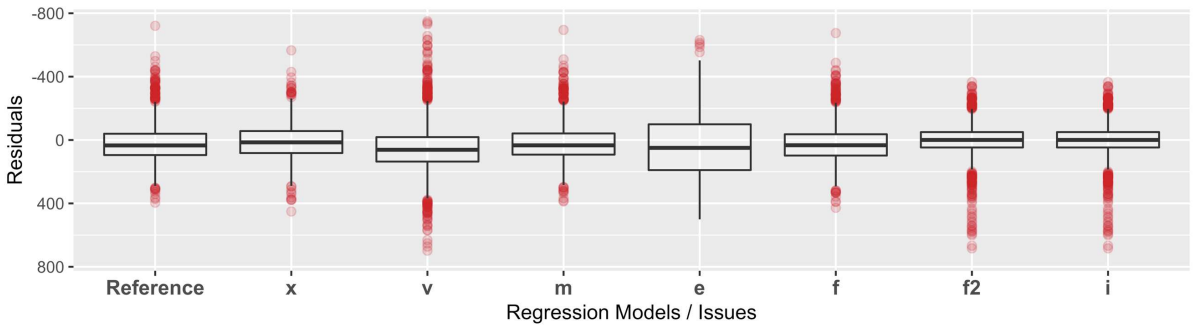


Figure 4: Boxplot distributions of the regression residuals ($y - y^*$) from the six issues and eight regressions (a to h). Outliers are given with a small transparency term.

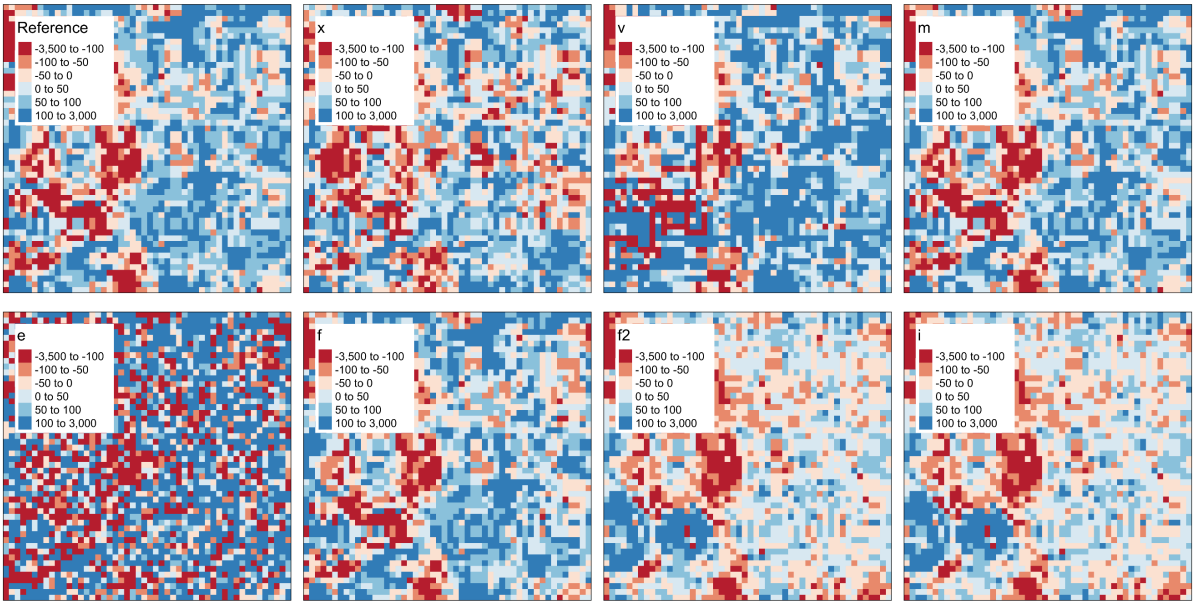


Figure 5: The residuals ($y - y^*$) arising from the six issues and eight regressions (a to h).

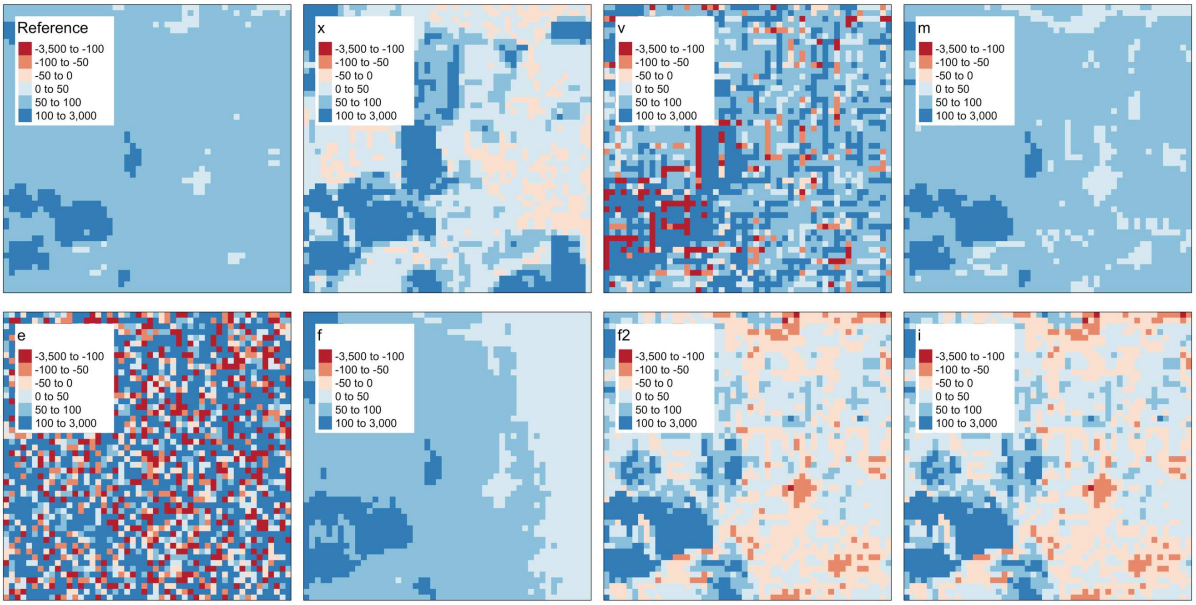


Figure 6: The differences between x_1 and the y^* values ($x_1 - y^*$) from the six issues and eight regressions (a to h).

Table 1. Model summaries for the eight regression analyses applied to the image case study.

Semantic issue and regression form		**Covariate**	**Estimate**	***t*-value**	**Pr. (>|*t*|)**
Reference: $y = f(x_1)$, multiple regression (MR)		x_1	0.943	548.022	0.000
***x* issue:** $y = f(x_1 + x_2)$, MR		x_1	1.193	126.506	0.000
		x_2	-0.283	-26.889	0.000
***v* issue:** $y = f(x_1)$, MR, x_1 aggregated to 60m		x_1	0.927	259.293	0.000
***m* issue:** $y = f(x_1)$, MR, square root of x_1		x_1	0.919	566.906	0.000
***e* issue:** $y = f(x_1)$, MR, random noise added to x_1		x_1	0.930	321.496	0.000
first ***f* issue:** $y = f(x_1)$, GWR*	*Min.*	-	0.919	298.183	0.000
	1st Quartile	-	0.935	431.140	0.000
	Median	x_1	0.943	484.059	0.000
	3rd Quartile	-	0.951	523.779	0.000
	Max.	-	0.992	572.124	0.000
second ***f* issue:** $y = f(x_1 + x_2)$, mixed model REML		x_1	0.524	28.783	0.000
		x_2	0.193	12.000	0.000
***i* issue:** $y = f(x_1 + x_2)$, mixed model ML		x_1	0.524	28.789	0.000
		x_2	0.193	11.997	0.000

*The GWR *t*-values are not corrected for multiple hypothesis tests (see Gollini et al., 2015).

From Figure 2, it is clear that for regression models (a, c, d, e and f) the y^* are much closer to x_1 than to y. This is particularly true for regressions (c and e), the ***v*** and ***e*** **issues,** respectively, where changes in the support of x_1 and the error associated with x_1 are very clearly reflected in y^*. For regression (d), the ***m*** **issue**, the y^* have a similar spatial pattern to x_1 as that found in the reference regression (a), but the y^* has reduced variability. This effect similarly results for regression (f), the first ***f*** **issue** with GWR, in comparison to the reference regression (a); but in this case, the y^* show the least spatial similarity to x_1, (at least for regressions using only a single covariate). Thus, accounting for spatial effects (the ***f*** **issue**) reduces the similarity between the y^* and x_1, as might be expected.

When an additional covariate, x_2 is added for the ***x*** **issue** (in regressions b, g and h), the y^* are no longer as close to x_1, but are now a reflection of x_1 and x_2, combined. Variability in y^* reduces when viewing the second ***f*** **issue** (a mixed model), in comparison to its reference regression (b); and this is similarly true with the first ***f*** **issue** (GWR), in comparison to its reference regression (a). On viewing only regressions (g and h), the chosen ***i*** **issue**, appears to have little to no effect on y^*, which was anticipated.

From Figure 3, the y^* from single covariate regressions (c, d, e and f, for issues ***v***, ***m***, ***e*** and ***f***, respectively) all display clear differences to that found with the reference regression (a). Similarly, the double covariate regression (b, for the ***x*** **issue**) provides clear differences in its y^* to that of the single covariate reference regression (a). Similarly, the double covariate regressions (g and h, for issues ***f*** and ***i,*** respectively), both display differences to that found with their reference regression (b), but not with each other. The results for Figure 4, mimic those given for Figure 3, but in terms of the residuals (y - y^*) rather than y^*.

From Figure 5, it is evident that regressions for five of the six issues (***x***, ***v***, ***m***, ***e***, and ***f***) all provide residuals that are different in spatial pattern to the reference regression and also to any alternative reference regression (i.e., the second ***f*** **issue** and the ***i*** **issue** with the ***x*** **issue** as their reference). For regressions representing the ***v*** and ***e*** **issues**, this effect can be quite striking. Figure 6, depicts how the spatial characteristics of y^* move closer to or further away from the spatial characteristics of the covariate x_1, depending on the illustrated semantic issue. Here, the regression for the ***m*** **issue** appears to provide y^* values that are the most similar to the covariate x_1, whilst the regressions for the ***v*** and ***e*** **issues** provide y^* values that are least similar.

The model inference summaries (Table 1) are only given for completeness and are largely uninteresting given that this is a synthetic case study. However, none of the six semantic issues result in a change of significance for any regression coefficients. The coefficient estimates can change, however; for example, the covariate x_2, negatively relates to y in the ***x*** **issue,** whereas the same relationship is a positive relationship in the second ***f*** **issue.**

4. Real case study: soils in the Loess Plateau, China

In this second case study, the predictive and inferential impacts of the ***x***, ***v***, ***m, f*** and ***i*** **issues** are illustrated for a soils dataset in the Loess Plateau, China. This area is dominated by thick loess deposits and suffers from intense erosion. It contains several fragile ecosystems (Chen et al., 2007; Wang et al., 2008) and has been the subject of much research, particularly focused on reducing the impacts of soil erosion by vegetation restoration, re-greening and sustainable agriculture initiatives (Fu et al., 2018; Luo et al., 2018; Hu et al., 2017; Li et al., 2017). Soil moisture is driven by limited precipitation, most of which falls between June and September. Understanding hydrological processes, and particularly soil conductivity, is critical in the Loess Plateau as these underpin the effectiveness of vegetation restoration

activities. Soil data were collected at 243 locations in the Loess Plateau, China (Figure 7) and analysed in the laboratory to generate measurements of several soil characteristics including soil physical properties (soil hydraulic conductivity, Ks, in cm min^{-1} and saturated soil water content, SSWC) and soil composition (% Clay, % Silt and % Sand). All soil measures were recorded at three different depths of 0 to 10 cm, 10 to 20 cm and 20 to 40 cm.

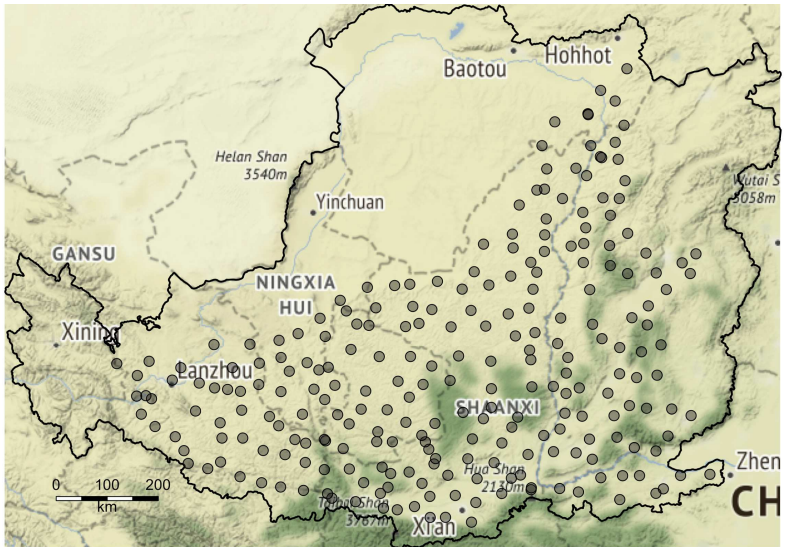


Figure 7. The 243 sample locations in the Loess Plateau, China, with a slight shading transparency to show sites that overlap.

In this study, Ks at 0 to 10 cm depth is considered as the target variable, *y*, with covariates SSWC at 0 to 10 cm and % Clay and % Silt both at 0 to 10 cm (% Sand at 0 to 10 cm is dropped to deal in a simple way with the compositional nature of the data). Additional covariates in terms of landscape (terrain) indices were obtained for each sample location from a series of DEMs at spatial resolutions of 30 m and 90 m. The 'terrain' function in the R 'raster' package (v2.6-7, Hijmans and van Etten, 2017) was used to generate measures of aspect (0 to 360°) and stream flow direction (encoded in powers of 2 to indicate compass direction). The aspect variable was transformed to East-ness by taking the cosine of the angles, and flow direction was converted to a measure of Flow Northeast-ness in a similar way using the sine function only.

The soil and terrain covariates allow us to demonstrate the ***x* issue,** and as the two terrain covariates (East-ness, Flow Northeast-ness) are calculated over different supports of 30 and 90 m, the ***v* issue,** also. The ***m* issue** is not present in this data, where it would have been ideal to have covariate data from different methods of measuring the exact same property using

three competing analytical techniques. However, we do have measurements of the *same* soil property collected at three different depths and, therefore, we use the SSWC at 20 to 40 cm, as alternative measurements of SSWC at 0 to 10 cm. This serves for illustration. If we had chosen to integrate the depth-dependent soils data, for example to 0 to 20 cm or 0 to 40 cm then we would be provided with an alternative ***v* issue.**

The nature of the alternative covariates for the study of the ***v* issue** and ***m* issue** are shown in Figures 8 and 9, respectively. The prediction accuracy of Ks at 0 to 10 cm should intuitively reduce for a regression calibrated with terrain variables over the larger supports (***v* issue**), and similarly, reduce for a regression calibrated with the deeper values of % SWCC (***m* issue**). These covariate decisions illustrate the potential for variation in regression outputs (predictions, coefficients and uncertainties) due to decisions over choice of support (Figure 8) and of measurement (Figure 9). The final data-driven semantic issue, the ***e* issue**, is not illustrated.

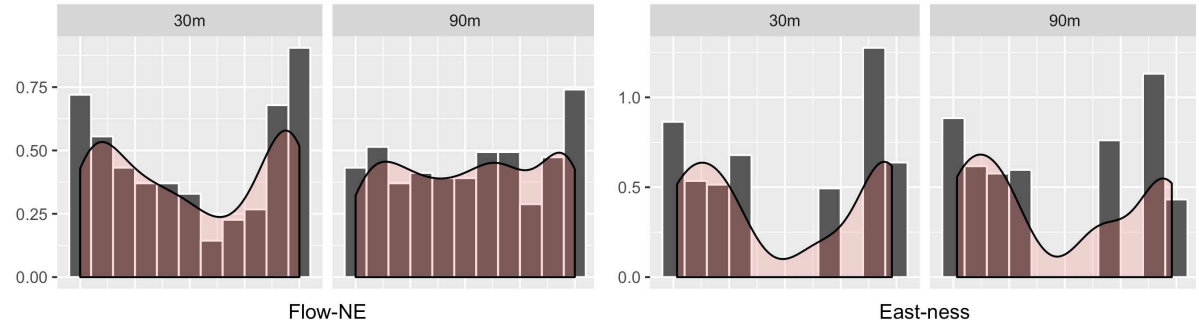


Figure 8. The ***v* issue** - variation in the terrain indices derived from different DEMs (probability densities). Note that the variance should decrease as sample support size increases.

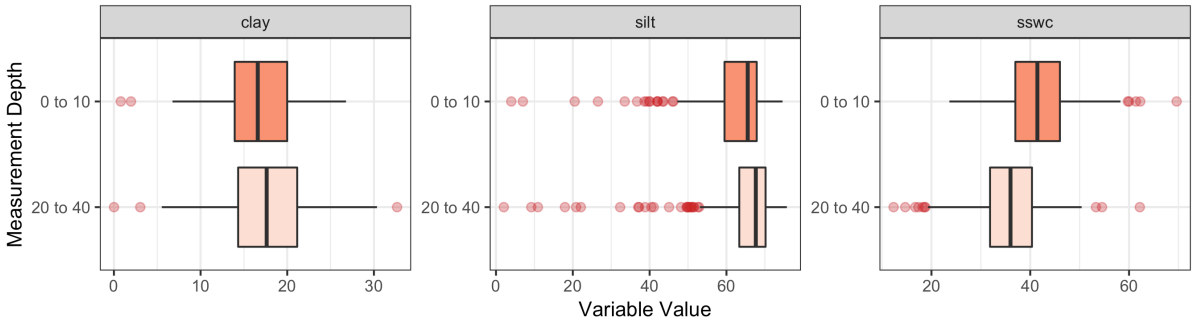


Figure 9. The contrived ***m* issue** - variation in the distributions of the soil property variables (% Clay, % Silt, % SSWC) sampled at the top and bottom depths (0 to 10 cm and 20 to 40 cm). The ***m* issue** is illustrated only using % SSWC and not % Clay and % Silt, also.

To illustrate the two model-driven semantic issues (the ***f*** and ***i*** **issues**), multiple regression, GWR and mixed models fitted using REML and ML were again specified. Thus, again two ***f*** **issues** are illustrated, where the ***i*** **issue** corresponds directly to the second ***f*** **issue** (via the mixed models). Again, the second ***f*** **issue** and its ***i*** **issue** are also given in the form used for the ***x*** **issue** and not the reference regression. The resultant seven regression analyses chosen to illustrate the ***x***, ***v***, ***m, f*** and ***i*** **issues** are as follows (with the differences to the reference regression highlighted in bold):

a) **Reference:** Ks at 0 to 10 cm = *f* (SSWC at 0 to 10 cm, Flow-NE and East-ness both at 30 m) with multiple regression;
b) ***x*** **issue:** Ks at 0 to 10 cm = *f* (SSWC at 0 to 10 cm, Flow-NE and East-ness both at 30 m; **Clay and Silt both at 0 to 10 cm**) with multiple regression;
c) ***v*** **issue:** Ks at 0 to 10 cm = *f* (SSWC at 0-10 cm, **Flow-NE and East-ness both at 90 m**) with multiple regression;
d) ***m*** **issue:** Ks at 0 to 10 cm = *f* (**SSWC at 20 to 40cm,** Flow-NE and East-ness both at 30 m) with multiple regression;
e) first ***f*** **issue:** Ks at 0 to 10 cm = *f* (SSWC at 0 to 10 cm, Flow-NE and East-ness both at 30 m) with **GWR;**
f) second ***f*** **issue:** Ks at 0-10 cm = *f* (SSWC at 0-10 cm, Flow-NE and East-ness both at 30 m; **Clay and Silt both at 0-10 cm**) with **mixed model fitted by REML;**
g) ***i*** **issue:** Ks at 0-10 cm = *f* (SSWC at 0-10 cm, Flow-NE and East-ness both at 30 m; **Clay and Silt both at 0-10 cm**) with **mixed model fitted by ML.**

Figure 10 maps the y^*'s. Model inference outputs for each of the seven regressions are summarised in Table 2.

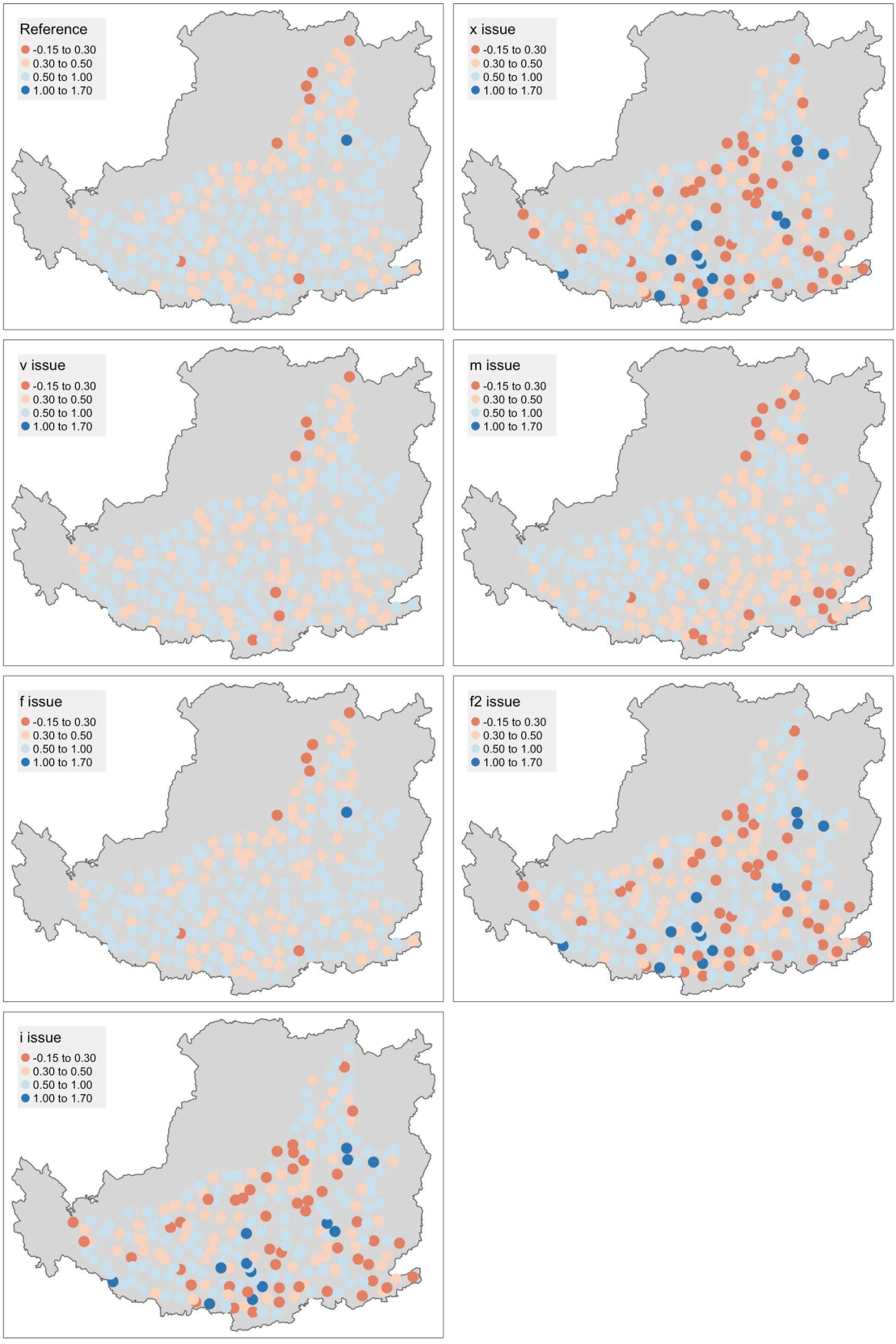


Figure 10. Maps of y^* arising from the five issues and seven regressions (a to g).

Table 2. Model summaries from the five issues and the seven regression analyses applied to the soils case study (changes from the reference regression highlighted in bold).

Regression form	**Covariate**	**Estimate**	***t*-value**	**Pr. (>|*t*|)**
Reference (multiple regression)	*SSWC at 0-10 cm*	0.013	22.292	0.000
	Flow-NE at 30m	0.085	2.660	0.008
	East-ness at 30m	-0.082	-2.398	0.017
***x* issue** (multiple regression)	*SSWC at 0-10 cm*	0.038	13.497	0.000
	Flow-NE at 30m	0.058	2.144	0.033
	East-ness at 30m	-0.074	-2.571	0.011
	Clay at 0-10 cm	0.047	6.755	0.000
	Silt both at 0-10 cm	-0.029	-9.837	0.000
***v* issue** (multiple regression)	*SSWC at 0-10 cm*	0.013	22.116	0.000
	Flow-NE at 90m	0.093	2.810	0.005
	East-ness at 90m	0.026	0.691	0.490
***m* issue** (multiple regression)	***SSWC at 20-40 cm***	0.014	19.194	0.000
	Flow-NE at 30m	0.096	2.733	0.007
	East-ness at 30m	-0.088	-2.345	0.020
first ***f* issue** (**GWR***, showing IQRs only)	*SSWC at 0-10 cm*	0.002	2.799	0.000
	Flow-NE at 30m	0.095	1.798	0.503
	East-ness at 30m	0.092	1.783	0.354
second ***f* issue** (**mixed model REML**)	*SSWC at 0-10 cm*	0.038	13.430	0.000
	Flow-NE at 30m	0.060	2.260	0.025
	East-ness at 30m	-0.080	-2.831	0.005
	Clay at 0-10 cm	0.044	6.185	0.000
	Silt both at 0-10 cm	-0.029	-9.488	0.000
***i* issue** (**mixed model ML**)	*SSWC at 0-10 cm*	0.038	13.430	0.000
	Flow-NE at 30m	0.060	2.254	0.025
	East-ness at 30m	-0.080	-2.815	0.005
	Clay at 0-10 cm	0.044	6.221	0.000
	Silt both at 0-10 cm	-0.029	-9.507	0.000

*The GWR *t*-values are not corrected for multiple hypothesis tests (see Gollini et al., 2015).

From Figure 10, it is useful to start with regression models (a, c, d and e) where the Ks (at 0 to 10 cm) predictions y^* are informed by SSWC, Flow-NE and East-ness covariates, and relate to the ***v***, ***m*** and (first) ***f* issues.** The regressions for (c, d and e) provide Ks prediction maps with a broadly similar spatial pattern to that found with the reference regression (a). However, the ***v* issue** (regression c) results in a reduced range of y^* and a likely reduction in its accuracy, and the possibility of missing areas of high or low Ks, that may be vital to the understanding of the soils process and its erosion. The ***m* issue** (regression d) appears to result in an increase in lower values of y^*, suggesting an under-prediction bias for Ks, especially towards the edges of the sampled area. Again, clear spatial interpretation issues for Ks may

result. For the ***f* issue** care must be taken, as unlike the ***v*** and ***m* issues,** where likely inaccuracies are expected (including their direction), it is not known which of multiple regression and GWR is the most accurate. GWR provides a cluster of unusually high Ks values to the mid-south of the sample area. If, as is likely, GWR predicts more accurately than multiple regression, then the mis-identification of such an area of high Ks values is likely to be problematic, especially for this region's soil health (a region of approximately 100 km by 200 km).

Conversely the maps of y^* are highly similar for regressions (b, f and g) where y^* are additionally informed by Clay and Silt covariates, encompassing the ***x*,** (second) ***f*** and ***i*** **issues**. This suggests a multiple regression for predicting Ks to be as worthy as a mixed model (i.e., the ***f* issue**), regardless of its identification method (i.e., the ***i* issue**). For the ***x*** **issue**, the y^* from regression (a) need to be compared to y^* from regression (b), where the addition of the Clay and Silt covariates has an obvious influence on y^*. As previously discussed, this outcome may relate to where missing covariates are reflected in spatial effects (in this case the non-stationary relationships of regression e), that may not be present if the regression were better-specified (i.e., with the extra covariates). However, this would need to be tested via a second GWR fit using all five covariates.

The model inference summaries are given in Table 2, and similarly indicate little difference in regressions (b, f and g), but clear differences for regressions (a, c, d and e). For the latter, this includes changes in coefficient sign (for East-ness) and changes in coefficient significance (for East-ness with regression c, and Flow-NE and East-ness with regression e). Again, the ***x* issue** in covariate choice, appears hugely important to the interpretation of this study's soils process.

5. Conclusions

This study illustrated how the semantics associated with regression prediction and inference are functions of choices over covariate data and regression models. The choice of covariates, the nature of their supports, the way that they are measured and their (unknown) errors were all shown to have an impact on the resultant prediction, and thereby its semantics. Prediction was also shown to be affected by model choices, specifically model specification and the

method of parameter identification. The implications of each of the issues identified are summarised in Table 3 and, of course, in any analysis several issues may be present and may interact. Many of these semantic effects are known but they are infrequently articulated and almost never routinely described in reports and scientific papers despite being the subject of wider study. Given the ubiquity of the use of regression within quantitative geography and the profound nature of the semantic impacts, this study presents an opportunity for researchers in quantitative geography to pause and update their thinking. In the extreme case of simple regression with one covariate x_1, the prediction is nothing more than a scaled version of that covariate, inheriting the same (scaled) spatial correlation and it is fundamentally naive to ignore this. Even small differences in data and model decisions can result in profound differences in outcomes including predictions, residuals, coefficient estimates, and the relationship between covariates and predictions. There is a wider and potentially more pressing context for these observations. Some advocates of machine learning have suggested that most of the issues identified in this paper can be ignored. However, as Marcus (2018, p15) states "deep learning is just a statistical technique, and all statistical techniques suffer from deviation from their assumptions".

Table 3. Implications of the different semantic issues

Issue	**Implications**
***x* issue** (which covariates to include in the model)	This limits the possibility space of y^*, which in turn, restricts its semantic meaning.
***v* issue** (support effects / measurement scales)	Choices (e.g., size, geometry, orientation, position) can result in different integrals x_v on the supports v which, in turn, can have a key effect on y^*.
***m* issue** (how x is measured)	The impact of any covariate x is dependent on its method of measurement and y^* will vary according to which form of x is specified.
***e* issue** (accuracy of measures of x)	The error e_x in x will propagate through the model transform to the prediction of y (related to the measurement of x above) and will have its own distribution and spatio-temporal character.
***f* issue (**specification of the model)	A semantic effect on y^* through the resultant linear weighted combinations of x's. Models for spatial data can vary greatly (e.g., linear vs. non-linear, stationary vs. non-stationary). The nature of the model parameters effect y^*.
***i* issue** (the statistical method of estimating the model parameters)	A semantic effect on y^* will be evident in the linear weights for combining the x's. Although often small, the choice of estimation or inference method for identifying the model may affect y^*.
***y* issue** (variation in y^* due to the specification of y)	The definition and sampling framework of y itself has a semantic effect on y^* at locations other than y (i.e., predictions at un-sampled locations).

The inclusion of GWR as a model choice raised some interesting issues in terms of the spatial structure of the predictions and residuals. For multiple regression, the impact on the predictions is a simple linear additive and multiplicative factor applied to the covariates. In GWR, spatial correlation is induced in the predictions that is not present in the covariates. In a mixed model, the impact includes spatial smoothing that is similarly not present in the covariates. More profoundly, as the nature of model information and model fit changes from location to location with GWR, this suggests that the semantics of y^* vary locally, also. This is evident in extensions to GWR, such as the geographically weighted LASSO (Wheeler, 2009) and geographically weighted elastic net logistic regression (Comber and Harris, 2018), where covariate selection is local; locally-compensated ridge GWR, where the method of estimation is local (Brunsdon et al., 2012); and autoregressive GWR (Brunsdon et al., 1998; Geniaux and Martinetti, 2017), where the method of identification is local.

In summary, the semantic issues highlighted in this paper have clear relevance to the many geographical studies that routinely use regression in one form or another, many of which inform the core challenges that the world is facing today, such as the global disease burden, global mortality, poverty, food insecurity, losses in biodiversity, and increasing pollution through urbanization. It is important that researchers think carefully about the semantics associated with any interpretation of predictions and the potential impact of semantic issues, given the large number of ways in which regression prediction may be affected by data and model choices. It is hoped that this study will raise awareness amongst quantitative geographers, challenge naive application of regression, lead to proper consideration of the semantic interpretation of regression predictions and introduce the semantic lens to data and model choices in regression modelling. This study supports this purpose by providing a clear conceptual framework that defines the multiple ways in which data and model choices can impact on prediction and inference. The uncertainty in prediction arising from these choices was deliberately omitted in establishing the core message of the paper but will be the subject of a future study.

Acknowledgements

This research was supported by the China-UK bilateral collaborative research on critical zone science (the Natural Environment Research Council Newton Fund NE/N007433/1, the National Natural Science Foundation of China NO. 41571130083) and the National Key

Research and Development Program of China (No. 2016YFC0501601). All of the data preparation, analyses and mappings were undertaken in R 3.5.1, the open source software. The code and data used in this analysis are available from https://github.com/lexcomber/SematicsOfStats. The DOI for the data and code are at https://zenodo.org/badge/latestdoi/123798087.

References

Anselin, L., 1990. Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science*, *30*(2), pp.185-207.

Anselin, L. (2010). "Thirty years of spatial econometrics." *Papers in regional science* 89 (1), 3-25.

Anyamba, A., Linthicum, K.J. and Tucker, C.J., 2001. Climate-disease connections: Rift Valley fever in Kenya. *Cadernos de saude publica*, *17*, pp.S133-S140.

Armstrong, M. (1984). "Problems with Universal Kriging." *Mathematieal Geology* 16, 101-108.

Atkinson, P. M., and Tate, N. J. (2000). Spatial scale problems and geostatistical solutions: a review. *Professional Geographer* **52,** 607-623. Basile et al., 2014;

Basile, R., M. Durban, R. Minguez, J.M. Montero, and J. Mur. (2014). "Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and non-linearities." *Journal of Economic Dynamics and Control* 48, 229-245.

Beck, L.R., Lobitz, B.M. and Wood, B.L., 2000. Remote sensing and human health: new sensors and new opportunities. *Emerging infectious diseases*, *6*(3), p.217.

Brunsdon, C., A.S. Fotheringham, and M. Charlton (1996). "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity." *Geographical Analysis,* 28, 281-298.

Brunsdon, C., A.S. Fotheringham, and M. Charlton. (1998). "Spatial nonstationarity and autoregressive models." *Environment and Planning A,* 30(6), 957-993.

Brunsdon, C., Charlton, M. and Harris, P. (2012). Living with Collinearity in Local Regression Models. In *Proceedings of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Brazil.

Carroll R J, Rupert D, 1988 Transformation and Weighting in Regression (Chapman and Hall, New York)

Casetti E, 1972, Generating models by the expansion method: applications to geographic research Geographical Analysis 4 81-91.

Chen, L.D., Wei, W., Fu, B.J., Lu, Y.H., 2007. Soil and water conservation on the Loess Plateau in China: review and perspective. Prog. Phys. Geogr. 31, 389–403.

Chilès JP, Delfiner P (1999) *Geostatistics - modelling spatial uncertainty*. Wiley, New York.

Christensen WF (2011) Filtered Kriging for Spatial Data with Heterogeneous Measurement Error Variances. Biometrika.

Comber A and Harris P (2018). Geographically weighted elastic net logistic regression. Paper accepted for publication in *Journal of Geographical Systems*, 20(4), 317-341.

Comber, A.J., Fisher, P.F., Wadsworth, R.A., (2005). What is land cover? *Environment and Planning B: Planning and Design*, 32:199-209.

Cressie, N. (1993). *Statistics for spatial data*. New York, USA: John Wiley and Sons.

Cressie, N. and N.H. Chan. (1989). "Spatial Modeling of Regional Variables." *Journal of the American Statistical Association,* 84, 393-401.

Derungs C, Wartmann F, Purves RS, and Mark DM. The meanings of the generic parts of toponyms: use and limitations of gazetteers in studies of landscape terms. Spatial Information Theory LNCS. 2013; 261-278. doi:10.1016/j.isprsjprs.2016.06.014.

Dong, G., R. Harris, K. Jones, and J. Yu. (2015). "Multilevel Modelling with Spatial Interaction Effects with Application to an Emerging Land Market in Beijing, China." PLoS ONE, 10(6), 1-18.

Frank IE, Friedman JH (1993). A Statistical View of Some Chemometrics Regression Tools." Technometrics, 35(2), 109{135.

Fu W, Lü Y, Harris P, Comber A and Wu L (2018). Peri-urbanization may vary with vegetation restoration: A large scale regional analysis. *Urban Forestry and Urban Greening*, 29: 77-87.

Geniaux, G., and D. Martinetti. (2017) "A new method for dealing simultaneously with spatial autocorrelation and spatial heterogeneity in regression models" *Regional Science and Urban Economics* https://doi.org/10.1016/j.regsciurbeco.2017.04.001.

Gollini, I., Lu, B., Charlton, M., Brunsdon, C. and Harris, P. (2015) GWmodel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models. *Journal of Statistical Software,* 63, 1-50.

Gotway, C. A., and L. J. Young. (2002). "Combining incompatible Spatial Data." *Journal of the American Statistical Association* 97(458), 632-48.

Harris P (2018) “A simulation study on specifying a regression model for spatial data - choosing between autocorrelation and heterogeneity effects” *Geographical Analysis*

Harvey, F. (2000). The social construction of geographical information systems. *International Journal of Geographical Information Science*, *14*(8), 711-713.

Hay, S.I., Guerra, C.A., Tatem, A.J., Atkinson, P.M. and Snow, R.W., 2005. Opinion—tropical infectious diseases: Urbanization, malaria transmission and disease burden in Africa. *Nature Reviews Microbiology*, *3*(1), p.81.

Hijmans, R.J. and van Etten, J., 2017. raster: Geographic data analysis and modeling. *R package version*, *2.6-7* https://cran.r-project.org/web/packages/raster/raster.pdf

Hu, J, Lü Y, Fu B, Comber AJ and Harris P (2017). Quantifying the effect of ecological restoration on runoff and sediment yields: A meta-analysis for the Loess Plateau. *Progress in Physical Geography*, DOI: 10.1177/0309133317738710

Hurvich CM, Simonoff JS, Tsai CL (1998). “Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion.” Journal of the Royal Statistical Society B, 60(2), 271–293.

Li T, Lü Y, Fu B, Comber AJ, Harris P and Wu L (2017). Gauging policy-driven large-scale vegetation restoration programmes under a changing environment: their effectiveness and socioeconomic relationships. *Science of the Total Environment*, 607-608: 911-919, DOI: 10.1016/j.scitotenv.2017.07.044

Liu, Y., Hu, J., Snell-Feikema, I., VanBemmel, M.S., Lamsal, A. and Wimberly, M.C., 2015. Software to facilitate remote sensing data access for disease early warning systems. *Environmental Modelling & Software*, *74*, pp.247-257.

Luo Y, Lü Y, Fu B, Harris P, Wu L and Comber A (2018). When multi-functional landscape meets critical zone science: advancing multi-disciplinary research for sustainable human well-being. *National Science Review*, https://doi.org/10.1093/nsr/nwy003

Marcus, G., 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Mark DM and Turk AG. Landscape categories in yindjibarndi: Ontology, environment, and language In Kuhn W, Worboys MF, Timpf S, editors. COSIT 2003 LNCS. 2003; 2825: 28–45.

Meloun, M., Militký, J., Hill, M., and Brereton, R.G. 2002. Crucial problems in regression modelling and their solutions. Analyst, 127(4), 433-450.

Murakami, D., and M. Tsutsumi. (2015). “Area-to-point parameter estimation with geographically weighted regression.” *Journal of Geographical Systems*, 17 (3), 207-225.

Openshaw S (1984) The Modifiable Areal Unit Problem. Geobooks, Norwich, England.

Pires P. Geospatial conceptualisation: a cross-cultural analysis on Portuguese and American geographical categorisations In Spaccapietra S, editor. Journal on Data Semantics III, LNCS. 2005; 3534: 196–212.

Robbins P. Fixed categories in a portable landscape: the causes and consequences of land-cover categorization. Env Plan A. 2001;33: 161–179.

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models ny using integrated nested Laplace approximations. Journal of the Royal Statistical Society B 71(2) 319-392

Smith B and Mark DM. Do mountains exist? Towards an ontology of landforms. Environ Plan B. 2003;30: 411–427.

Smith, B. (2001). Fiat objects. *Topoi*, *20*(2), 131-148.

Smith, B., & Varzi, A. C. (2000). Fiat and bona fide boundaries. *Philosophical and Phenomenological Research*, 401-420.

Turk AG, Mark DM and Stea D. Ethnophysiography. In: Mark DM, Turk AG, Burenhult N, Stea D, editors. Landscape in language Transdisciplinary perspectives. Amsterdam: John Benjamins Publishing; 2011. pp 25–45.

Wang J, Stein A, Gao B, Ge Y (2012) A review of spatial sampling. Spat Stat 2:1–14Wang, L., Wang, Q., Wei, S., Shao, M.A., Li, Y., 2008. Soil desiccation for Loess soils on natural and regrown areas. For. Ecol. Manag. 255, 2467–2477.

Welham S, Cullis B, Gogel B, Gilmour A, Thompson R. 2004. Prediction in linear mixed models. *Australian and New Zealand Journal of Statistics* **44**: 571–586.

Wheeler, D 2009. Simultaneous Coefficient Penalization and Model Selection in Geographically Weighted Regression: the Geographically Weighted Lasso. *Environment and Planning A*, 41(3), 722–742.

Zhang, J., P.M. Atkinson, and M. Goodchild (2014). *Scale in Spatial Information and Analysis*. CRC Press.

Zou H and Hastie T, (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2): 301–320