Semantic considerations of local statistical modelling: what does it all *mean*?

Alexis Comber[*1], Peter Atkinson[2], Paul Harris[3]

[1]School of Geography, University of Leeds, Leeds, LS2 9JT, UK
[2] Faculty of Science and Technology, Lancaster University, Lancaster LA1 4YW, UK
[3]Rothamsted Research, North Wyke, Okehampton, Devon, EX20 2SB, UK
*Email: a.comber@leeds.ac.uk

Abstract

This paper explores the spatially varying semantics associated with local statistical models. As demonstration, it applies two different geographically weighted (GW) models to two case studies and shows how the characteristics and thus the semantics of the factors associated with house price in Hanoi, or those associated with Trump supporting counties in the recent US election, both vary with location. The paper addresses a key gap in the consideration of many geocomputational models concerning the meaning of their outputs and what their outputs represent. The paper starts to develop these arguments in relation to the underlying assumptions of statistical models - in this case linear regression and linear discriminant analyses and their corresponding localised versions.

Keywords: cognitive spatial information theory, representation.

1. Introduction

Geography has long considered issues associated with how geographic features (objects, processes phenomena, etc.) are conceptualised, measured and ultimately represented in spatial databases. It has resulted in theoretical developments in cognitive spatial information science (e.g. Smith and Mark, 2003; Pires 2005; Mark and Turk 2003; Turk et al 2011; Derungs et al 2013; Robbins 2001). Semantics are at the core of this corpus of research. Semantics describe what features encoded in data *mean*, the landscape perceptions they embody, the *weltanschauung*, underlying ontologies and epistemologies of measurement. Semantics in geography (*cf* semantics in computer science) is concerned with *how* we represent the real world in our data (and maps!). It seeks to elicit deeper understandings of the *meaning* of database objects and what they represent. This is 'core Geography'.

Much of the research into geo-spatial semantics has focused on the *meaning* of spatial information as related to measurement and representation (e.g. Goodchild et al, 2007; Frank, 2003) and not the semantic aspects associated with many of the algorithmic developments in geocomputation and spatial analysis. Despite much work in GIScience, as yet very little consideration has been given to the semantic implications of geo-statistical methods like those found within the geographically weighted (GW) framework (Gollini et al. 2015), or has reflected on the semantic nature of such localised model outputs. This paper addresses this critical gap by exploring GW regression (GWR) (Brunsdon et al, 1996) and GW discriminant analysis (GWDA) (Brunsdon et al. 2007), each of which seek to replace whole map statistics with ones that are geographically and locally more sensitive. Other approaches to local modelling exist (Griffith 2008; Fouedjio 2016) and GW models are explored to illustrate the problem.

One of the implications of local (rather than global) statistical models is that the semantics of the model outputs – what the model means – also varies locally. Through two case studies this papers demonstrates the origins and implications of spatially varying semantics associated with local statistical models and considers what the implications are for the way that model results are interpreted.

2. Case Studies

2.1. Hanoi house price

A household survey undertaken in 2014 in Hanoi, Vietnam captured data on house price and associated variables describing socio-economics, local area perceptions and house characteristics. A global stepwise model selection resulted in 11 predictor variables as inputs to a predictive model of house price. A linear regression, fitted by ordinary least squares (OLS), was used to generate a global, non-spatial predictive model, and GWR was used to explore the spatial heterogeneity in the same predictive relationships. Coefficient estimates and their significance derived from the linear regression are shown in Table 1, together with the coefficient estimates from GWR. The latter indicate how factors associated with house price vary locally and these are mapped in Figure 1. The maps show how the *relationships* between house price and the predictor variables vary in different areas in Hanoi. The implication of these spatial variations is that the semantics of each predictive local house price model changes in different places.

	Linear regression				GWR		
Variable	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**1st Qu**	**Median**	**2nd Qu**
(Intercept)	4758.01	836.52	5.69	0.000	4634.66	6276.14	7288.28
Distance to city	-0.61	0.07	-8.55	0.000	-1.04	-0.81	-0.58
Ground floor area	10.52	2.05	5.14	0.000	3.60	11.38	19.51
Total floor area	14.14	1.76	8.05	0.000	12.15	15.81	23.76
No. residents	200.17	95.46	2.10	0.036	58.18	162.52	342.09
Income	44.15	15.84	2.79	0.005	-18.11	9.51	62.85
No. of storeys	-306.69	202.79	-1.51	0.131	-745.93	-342.44	-41.25
Car ownership (T)	1074.50	748.43	1.44	0.152	440.45	1295.94	1949.88
University (T)	1035.31	378.38	2.74	0.006	280.41	600.58	930.49
Mezzanine (T)	-977.60	549.35	-1.78	0.076	-1223.01	-784.02	-439.21
Shopfront (T)	960.40	378.24	2.54	0.011	383.45	870.36	1113.31
Residential (T)	-1209.98	417.60	-2.90	0.004	-1603.09	-976.37	-503.13

Table 1: The linear regression coefficient estimates and t-values, and summaries of GWR coefficient estimates only (T indicates True).

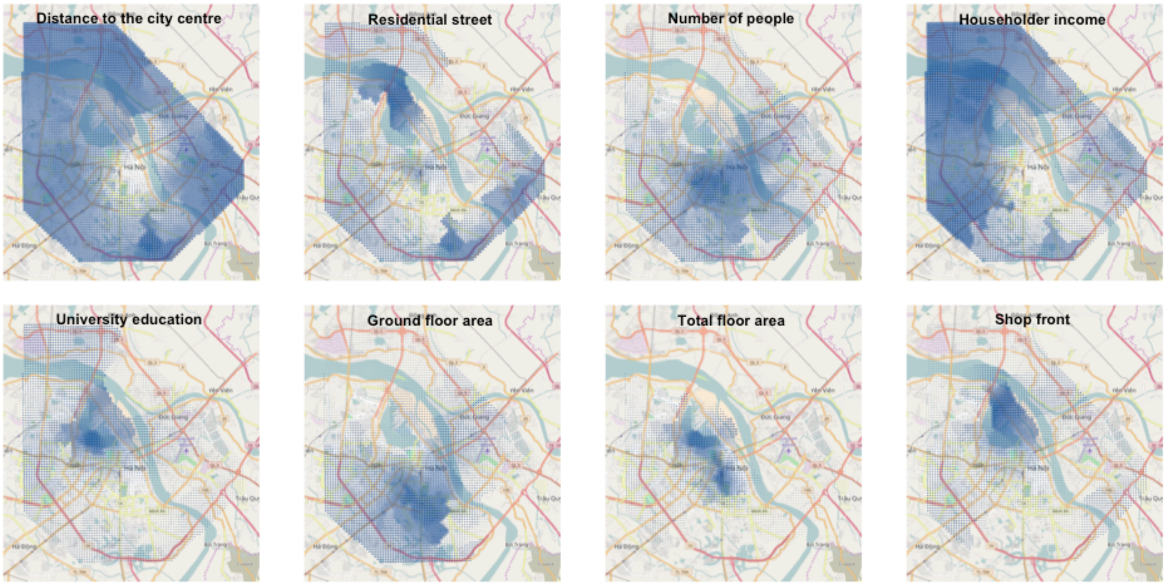

Figure 1: The spatial distribution of the GWR coefficient estimates with a transparency term and an Open Street Map backdrop. The plot character size indicates the coefficient value, rescaled to the range [0, 1].

2.2. Geo-demographic classification of Trump support

County level voting data from the recent presidential election in the USA was linked to census data describing 5 variables and used to model 3 classes of county: Trump supporting, Clinton supporting and Borderline. The voting data was downloaded from Tony McGovern Github site (https://github.com/tonmcg/County_Level_Election_Results_12-16). A discriminant analysis (DA) was used to classify each county to one of the 3 categories. A GWDA is a local adaption of DA, but where the discrimination rules are allowed to vary across space, with the mean and covariance estimates replaced with GW mean estimates and GW covariance estimates (Brunsdon et al, 2007; Lu et al, 2014). Summaries of the coefficients of the linear discriminants arising from the DA and GW DAs are shown in Table 2. *NB* the global coefficients are outside of the GWDA coefficient range, reflecting the way that discrimination rules are determined. Table 2 illustrates the spatially varying nature of GWDA approach.

Linear Discriminants	Global Coefficients of LDs	GW 1st Quartile	GW Median	GW 3rd Quartile
% Unemployed LD1	-0.051	-0.407	-0.287	-0.145
% College LD1	-0.118	-0.358	-0.273	-0.198
% Over 65 LD1	0.015	-0.157	-0.100	-0.058
% Urban LD1	-0.003	-0.023	-0.013	-0.005
% White LD1	0.057	0.074	0.087	0.097
% Unemployed LD2	-0.033	-0.437	-0.326	-0.237
% College LD2	0.068	-0.113	-0.018	0.217
% Over 65 LD2	-0.075	-0.348	-0.235	-0.104
% Urban LD2	0.010	-0.016	0.007	0.037
% White LD2	0.045	0.027	0.065	0.088

Table 2: The coefficients of the linear discriminants (LD1 and LD2) from a global and local DAs.

Figure 2 shows the spatial variation in the GWDA group means for the classes of ‘Trump supporting county’ in comparison to the corresponding global group mean. The GWDA results suggest, for example, that a Trump county in the Midwest is much more highly associated with the number of over 65s than in the West or Florida. It is evident that locally, the definition and semantics of a ‘Trump supporting county’ changes depending on location and different factors have different group mean values in different locations.

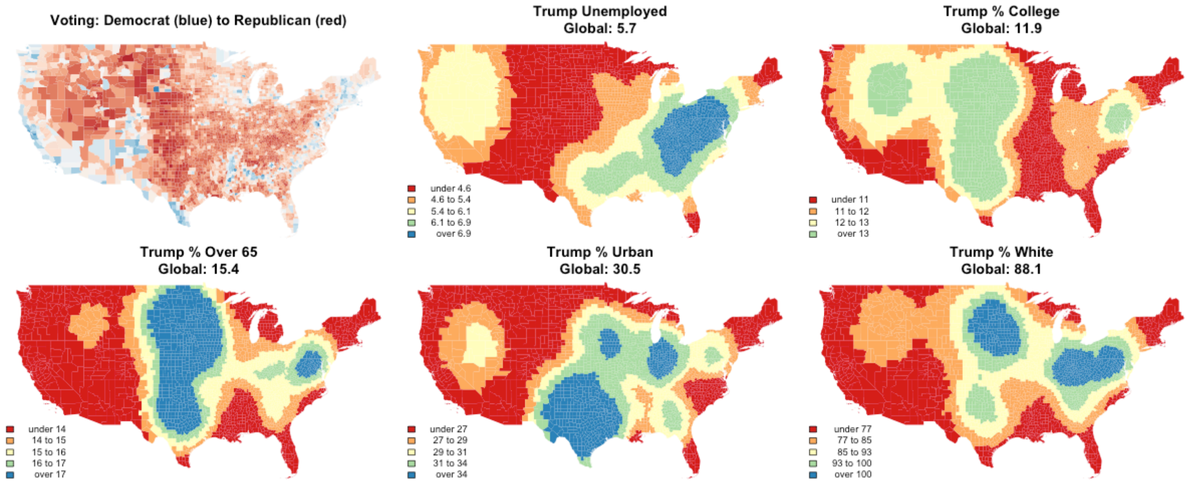

Figure 2: The spatial variation in the group means for the category of ‘Trump county’. High levels of difference are shown in red and blue.

3. Discussion

The case studies show how the different factors or variables (*X*'s) vary spatially in the degree to which they are associated with *Y*. Thus the semantics of *Y* vary spatially, suggesting that in Hanoi different socio-economic, infrastructural and house characteristics associated with property price vary spatially: high coefficients for car ownership in some areas and low in others; high variation in the impact on price houses being on residential streets; and variation in room number etc. In the US elections, the GWDA shows how the semantics of a ‘Trump support’ varies in different parts of the country, for example unemployment in the East, the over 65s in the Midwest, the link to the urban proportion in Texas. These suggest that the factors related to Trump support – and therefore the semantics of Trump support – are different in different places, and indicate how the different underlying socio-economic pressures played out in that support.

This suggests the need for deeper consideration of the nature of the construction of the localised models, much of which is a transference of known (but forgotten) issues found in the global case. Room precludes this discussion in this extended abstract, but will be expanded upon in the full paper arising from this work, and in the conference presentation. Semantic considerations on *Y* include, for example: that predicted *Y* (*Y**) is always a function of *X* and never *Y*; the effect of the support *v* of the *X* on *Y*,* where *X* can have different supports (or are misaligned); the consequences of model selection, model (mis-)specification and issues of model component identification on *Y**; and also issues surrounding the sampling framework itself on *Y*.*

To conclude, local models such as GWR or GWDA describe how processes and relationships very spatially. But the fundamental tenet of local approaches is that at each location where *Y* is being predicted the 'meaning' of *Y*, what *Y* represents – **the semantics *Y*** – changes at each location because local models or predictions of processes, relationships and outcomes, implicitly define *Y* differently in different places. The implications of this are that different factors will have different salience in different places, perhaps reflecting different local policies, but require different responses and solutions in different places.

4. Acknowledgements

This research was supported the Natural Environment Research Council Newton Fund grant (NE/N007433/1) and a UK Biotechnology and Biological Sciences Research Council grant (BBSRC BB/J004308/1).

4. References

Brunsdon C, Fotheringham AS, Charlton M (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28, 281–289.

Brunsdon, C., Fotheringham, S., & Charlton, M. (2007). Geographically weighted discriminant analysis. *Geographical Analysis*, *39*(4), 376-396.

Derungs C, Wartmann F, Purves RS, and Mark DM (2013). The meanings of the generic parts of toponyms: use and limitations of gazetteers in studies of landscape terms. *Spatial Information Theory LNCS*, 261-278.

Fouedjio, F. (2016). Second-order non-stationary modeling approaches for univariate geostatistical data. *Stochastic Environmental Research and Risk Assessment*, 1-20.

Frank, A. U. (2003). Ontology for spatio-temporal databases. In *Spatio-Temporal Databases* (pp. 9-77). Springer Berlin Heidelberg.

Goodchild MF, Yuan M and Cova TJ (2007) Towards a general theory of geographic representation in GIS, *International Journal of Geographical Information Science*, 21:3, 239-260

Griffith, D. A. (2008). Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A*, 40(11), 2751-2769.

Lu, B., Harris, P., Charlton, M., & Brunsdon, C. (2014). The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models. *Geo-spatial Information Science*, *17*(2), 85-101.

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

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

Robbins P (2001). Fixed categories in a portable landscape: the causes and consequences of land-cover categorization. *Environment and Planning A*, 33:161–179.

Smith B and Mark DM (2003). Do mountains exist? Towards an ontology of landforms. *Environment and Planning B*. 30: 411–427.

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