Specifying regression models for spatio-temporal data sets

Paul Harris*[1], Alexis Comber[2], Narumasa Tsutsumida[3]

[1]Rothamsted Research, North Wyke, Okehampton, Devon, EX20 2SB, UK
[2]School of Geography, University of Leeds, Leeds, LS2 9JT, UK
[3]Graduate School of Global Environmental Studies, Kyoto University, Kyoto 606-8501
*Email: paul.harris@rothamsted.ac.uk

Abstract

This study investigates regression model specification issues with respect to the consideration of autocorrelation effects and/or relationship heterogeneity effects with spatio-temporal panel data. Models include: simultaneous autoregressive regression, geographically weighted regression and their corresponding extensions to the space-time case. This study focuses on the exploration and specification of such models and associated identification issues, and how they relate to ultimate goals of prediction and inference. As a case study, 16 years of livestock (cattle) population data in Mongolia are informed by environmental, climatic, socio-economic and agricultural covariates.

Keywords: Autocorrelation, Relationship Heterogeneity, Identification, Misspecification.

1. Introduction

Often when choosing a regression to model spatio-temporal data, it is not clear what, if any, spatial and/or temporal dependencies or relationships should be incorporated. Decisions also depend on sample information and its configuration, both in space and in time. Given that this study's data is relatively information-rich in the spatial dimension (341 contiguous area units), but relatively information-poor in the temporal dimension (16 consecutive years), there are a distinct class of methodological approaches to choose from.

One approach is to consider each year in turn and implicitly assume temporal independence, with 16 separate regressions. If this is the case, model decisions can then be taken on whether, for example, the focus should be on regressions with spatial autocorrelation effects, say with a simultaneous autoregressive regression (SAR) (Anselin 1988), or regressions with spatial heterogeneity effects with respect to data relationships, say with a geographically weighted regression (GWR) (Brunsdon et al. 1996)? Furthermore, should we try to capture both effects, for example an autoregressive GWR model (GWR-SAR) (Brunsdon et al. 1998)? It also possible to link (Griffith 2008), or fuse these two spatial effects together (Assunção 2003) – regression models that are commonly specified within a Bayesian inferential framework. Furthermore, what should be done for spatial effects that are likely to vary at different spatial scales, which in turn can be from an error variance (e.g. Milne et al. 2010) or relationship viewpoint (Lu et al. 2017; Murakami et al. 2017)?

Difficulties arise in that it is not easy to determine which effects are more important, where autocorrelation and heterogeneity are commonly strongly interrelated causing problematic identification issues and analytical confounders (Armstrong 1984; Anselin 1990), which in turn, tend to depend on the spatial characteristics of the covariate data set. In some instances, it can be pragmatic to proceed with regressions with no spatial effects, but possibly with the coordinates as spatial covariates, say; or alternatively seek for additional (or re-observed) covariates such that any observed spatial effects are removed by their presence (Cressie and Chan 1989).

Of course assuming temporal independence is naïve, and as a result, each annual regression is likely to be missing valuable time-dependent information. However, all of the specification difficulties stated above, are now present in the spatio-temporal domain. For example, should we focus on autocorrelations effects, say with a spatial panel regression (SPR) (Anselin et al. 2008; Millo and Piras 2012), or focus on heterogeneity effects, say with a geographical and temporal weighted regression (GTWR) (e.g. Huang et al. 2010; Fotheringham et al. 2015). Further still should we investigate a hybrid of both models, say with a geographically weighted panel regression (GWPR) (Yu 2010) or an autoregressive GTWR hybrid, call this GTWR-SAR, say (Wu et al. 2014).

2. A case study: 16 years of cattle population data in Mongolia

For demonstration, we analyse annual cattle population data for 341 soums (second-level administrative units) in Mongolia from 1990-2006 (as described in Tsutsumida et al. 2017), which are considered a function of the following four covariates: (i) annual mean normalised difference vegetation index (NDVI), (ii) annual mean rainfall (Rain), (iii) the number of households working with livestock (NH) and (iv) the number of reported cattle losses (Loss).

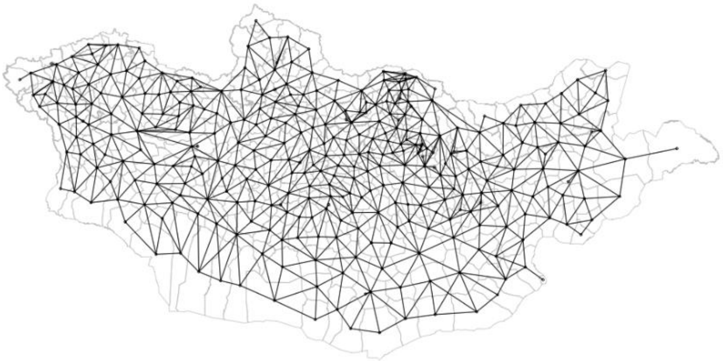
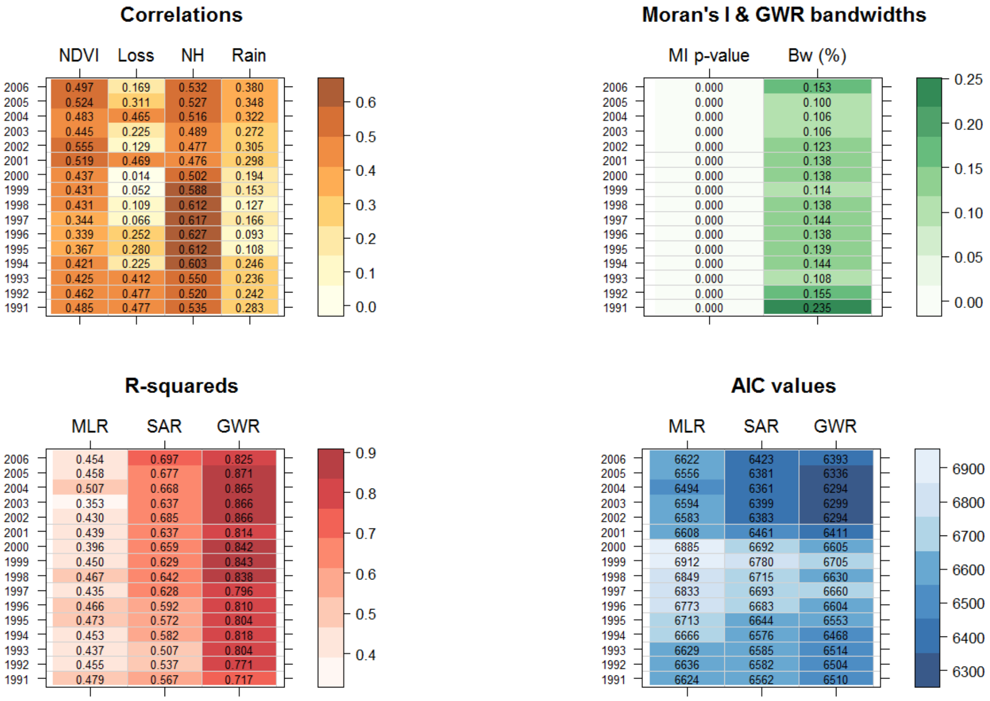


Figure 1: The soum adjacencies used to construct the neighbourhood weight matrix in the Moran's *I* test and the SAR model.

2.1. Spatial regressions

The first step in the analysis is to explore the data on a year by year basis for: (a) global correlations between the cattle (response) data and the four covariates; (b) significant spatial autocorrelation in the error term from a multiple linear regression (MLR) via Moran's *I*; (c) evidence of spatial heterogeneity in data relationships via an automatically found bandwidth for a GWR fit (using an AIC approach); (d) global R-squared values for MLR, SAR and GWR models; and (e) global AIC values from MLR, SAR and GWR models. Exploratory analyses (b) and (c) look for evidence that contradicts the underlying assumptions of the (non-spatial) MLR model - in that the errors are assumed independent and that data relationships are assumed fixed across space, respectively. R-squared values provide a handle of relative model fit with a prediction focus, whilst AIC values similarly account for model fit whilst penalising for model complexity.

MLR is fitted using ordinary least squares, the SAR model chosen for this study accounts for spatial autocorrelation via the error term and is fitted using maximum likelihood, whilst GWR is fitted via a weighted least squares procedure. The SAR model and the Moran's *I* test each account for spatial structure in the error term via the interactions between each pair of spatial units as captured by a spatial weights matrix derived from soum contiguity (Figure 1). Thus for MLR and SAR, single regressions are fitted, where the latter has an additional parameter which controls the degree of autocorrelation in the error term. Whereas for GWR, a series of local MLRs are calibrated at any location i with observations near to i given more influence than observations further away by weighting them via a distance-decay, kernel weighting function. The locations of the 341 observation points are taken as the centroids of the soums; and in this study, adaptive bi-square bandwidths are specified (given as a percentage).



Correlations

	NDVI	Loss	NH	Rain
2006	0.497	0.169	0.532	0.380
2005	0.524	0.311	0.527	0.348
2004	0.483	0.465	0.516	0.322
2003	0.445	0.225	0.489	0.272
2002	0.555	0.129	0.477	0.305
2001	0.519	0.469	0.476	0.298
2000	0.437	0.014	0.502	0.194
1999	0.431	0.052	0.588	0.153
1998	0.431	0.109	0.612	0.127
1997	0.344	0.066	0.617	0.166
1996	0.339	0.252	0.627	0.093
1995	0.367	0.280	0.612	0.108
1994	0.421	0.225	0.603	0.246
1993	0.425	0.412	0.550	0.236
1992	0.462	0.477	0.520	0.242
1991	0.485	0.477	0.535	0.283

Moran's I & GWR bandwidths

	MI p-value	Bw (%)
2006	0.000	0.153
2005	0.000	0.100
2004	0.000	0.106
2003	0.000	0.106
2002	0.000	0.123
2001	0.000	0.138
2000	0.000	0.138
1999	0.000	0.114
1998	0.000	0.138
1997	0.000	0.144
1996	0.000	0.138
1995	0.000	0.139
1994	0.000	0.144
1993	0.000	0.108
1992	0.000	0.155
1991	0.000	0.235

R-squareds

	MLR	SAR	GWR
2006	0.454	0.697	0.825
2005	0.458	0.677	0.871
2004	0.507	0.668	0.865
2003	0.353	0.637	0.866
2002	0.430	0.685	0.866
2001	0.439	0.637	0.814
2000	0.396	0.659	0.842
1999	0.450	0.629	0.843
1998	0.467	0.642	0.838
1997	0.435	0.628	0.796
1996	0.466	0.592	0.810
1995	0.473	0.572	0.804
1994	0.453	0.582	0.818
1993	0.437	0.507	0.804
1992	0.455	0.537	0.771
1991	0.479	0.567	0.717

AIC values

	MLR	SAR	GWR
2006	6622	6423	6393
2005	6556	6381	6336
2004	6494	6361	6294
2003	6594	6399	6299
2002	6583	6383	6294
2001	6608	6461	6411
2000	6885	6692	6605
1999	6912	6780	6705
1998	6849	6715	6630
1997	6833	6693	6660
1996	6773	6683	6604
1995	6713	6644	6553
1994	6666	6576	6468
1993	6629	6585	6514
1992	6636	6582	6504
1991	6624	6562	6510

Figure 2: Top: Global relationships with cattle population, Moran's *I* p-values for residuals from MLR fit, and GWR bandwidths. Bottom: R-squared and AIC values for MLR, SAR and GWR.

Figure 2 shows an emergent picture of spatial structure amongst the data across the 16 years. The global correlations, suggest reasonable predictive power of cattle population, especially for covariates NDVI and NH. GWR bandwidths are typically around 10 to 24% of nearby data suggesting that the same relationships may vary across space and are not fixed. For autocorrelation, all Moran's *I* p-values are highly significant. Thus both spatial heterogeneity and spatial autocorrelation appear viable options when it comes to specifying each year's regression model. Although, we don't actually know which spatial process best represents the yearly data, as these effects are strongly interrelated (Anselin 1990, p.204). The R-squared and AIC values clearly indicate that GWR will not only provide the best predictions of cattle population, but also provides the most parsimonious model.

Thus on balance, GWR appears the most promising model, although we have not as yet, investigated its inferential properties, where issues such as collinearity and multiple hypothesis testing also need to be assessed - issues that a SAR model is more easily adapted to cater for. Furthermore, to compete this initial year-on-year investigation, it may be of value to calibrate and fit hybrid GWR-SAR models, as proposed by Brunsdon et al. (1998).

2.2. Spatio-temporal regressions

The initial year-on-year investigations provide evidence for the suitability of both GWR and SAR to be extended to the space-time case, through GTWR and SPR models, respectively. We should also consider hybrids in GWPR and GTWR-SAR models. Outputs from all such spatial (fitted 16 times) and spatio-temporal models (fitted once only) can be assessed together, and Table 1 summarises their core properties. For inference, it is vital that temporal, spatial and spatio-temporal dependencies are accounted for, otherwise biased predictions, coefficient estimates and their standard errors will result. Thus each of the eight study models will tend to a certain bias according to their stated properties.

Model	Type	Fixed coefficients?	No. of coefficients	Autocorrelation parameters?	No. of autocor. parameters
MLR	spatial only	yes	5	no	0
SAR	spatial only	yes	5	yes	1
GWR	spatial only	no	341 x 5	no	0
GWR-SAR	spatial only	no	341 x 5	yes	341
SPR	spatio-temporal	yes	5	yes	1
GWPR	spatio-temporal	no	341 x 5	yes	341
GTWR	spatio-temporal	no	16 x 341 x 5	no	0
GTWR-SAR	spatio-temporal	no	16 x 341 x 5	yes	5456

Table 1: Summary of model properties when applied to the case study data.

So as to match the SAR model chosen for this study, which accounts for spatial autocorrelation via the error term, the GWR-SAR, SPR, GWPR and GTWR-SAR models are similarly specified. Thus we do not to consider lagged SAR-based models. SPR and GWPR use observations repeated over time, where in this case, the data are 'balanced' as the number of cross sectional observations is constant across the 16 time periods. Given space limitations, we only present results for 'in-sample' predictions of cattle population in 2006 for MLR, SAR, GWR and GWR-SAR (Figure 3). Unsurprisingly, GWR-SAR provides the best fit.

Observe that only the GTWR and GTWR-SAR models truly account for any temporal dependencies. In this respect, their specification can be highly involved where complications arise in that a suitable space-time metric is not immediately obvious. The concept of 'nearness' is also different in the temporal dimension to that intuitively assumed in the spatial dimension. Given these difficulties, details of the implementation of such models are presented in a sister paper to this one (Comber et al. 2017), but where the outputs of all eight study models are compared here.

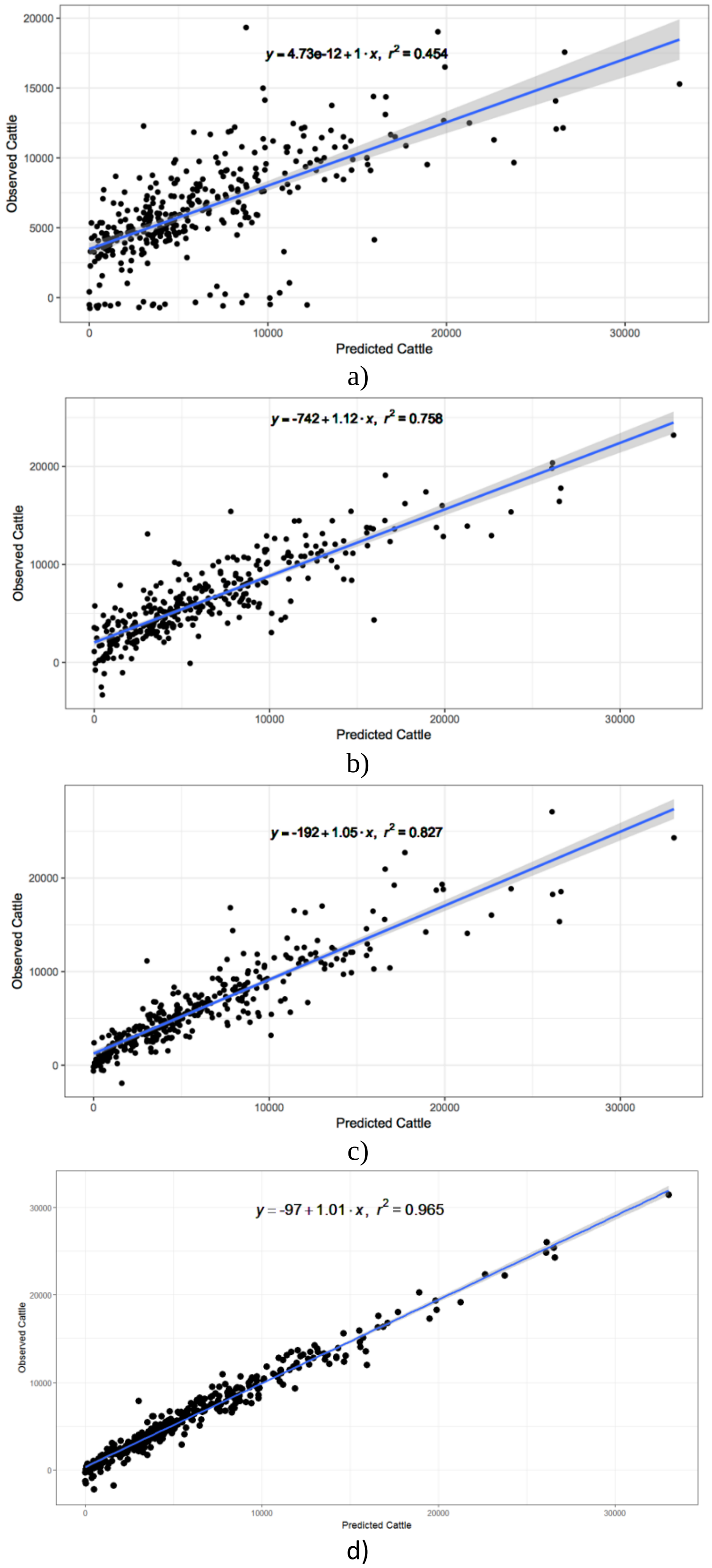


Figure 3: Fitted against observed values of Cattle in 2006 under a) the MLR, b) SAR, c) GWR, and d) GWR-SAR models.

3. Concluding remarks

This study aims to provide guidance on regression specification with respect to autocorrelation effects and/or relationship heterogeneity effects with spatial panel data. Eight models are considered, four are purely spatial that need to be repeated at each time interval, whilst four are spatio-temporal that consider the spatial panel data as a whole. Models stem from either a SAR or GWR paradigm.

4. Acknowledgements

This research was supported the Natural Environment Research Council Newton Fund grant (NE/N007433/1), a UK Biotechnology and Biological Sciences Research Council grant (BB/J004308/1), Sinfonica Statistical GIS Research Grants, JSPS KAKENHI (15K21086) and the KU SPIRITS project. The authors would also like to thank the National Statistics Office of Mongolia for providing the livestock population data.

5. References

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Anselin, L. (1990) Spatial Dependence and Spatial Stucture Instability in Applied Regression Analysis. *Journal of Regional Science*, 30, 185-207.

Anselin, L., Le Gallo, J., Jayet, H. (2008) Spatial Panel Econometrics." In L Matyas, P Sevestre (eds.), The Econometrics of Panel Data - Fundamentals and Recent Developments in Theory and Practice, pp. 624-660. Springer-Verlag.

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

Assunção R.M. (2003) Space varying coefficient models for small area data. *Environmetrics* 14, 453-473.

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

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

Comber, A., Harris, P., & Tsutsumida, N., (2017) Time: the late arrival at the Geocomputation party and the need for considered approaches to spatio-temporal analyses. Geocomputation 2017, Leeds.

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

Fotheringham, A. S., Crespo, R., & Yao, J. (2015). Geographical and temporal weighted regression (GTWR). *Geographical Analysis*, *47*(4), 431-452.

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

Huang, B., Wu, B., & Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, *24*(3), 383-401.

Lu, B., Brunsdon, C., Charlton, M. & Harris, P. (2017) Geographically Weighted Regression with Parameter-Specific Distance Metrics. *International Journal of Geographical Information Science 31(5), 982-998*

Millo, G. & Piras, G. (2012) splm: Spatial Panel Data Models in R. Journal of Statistical Software, 47 (1).

Milne, A.E., Webster, R., Lark, R.M. (2010). Spectral and wavelet analysis of gilgai patterns from air photography. Australian Journal of Soil Research, 48, 309–325.

Murakami, D., Yoshida, T., Seya, H., Griffith, D.A., & Yamagata, Y. (2017) A Moran coefficient-based mixed effects approach to investigate spatially varying relationships. *Spatial Statistics,* 19, 68-89.

Tsutsumida, N., Harris, P., & Comber, A. (2017) The application of a geographically weighted principal components analysis for exploring 23 years of goat population change across Mongolia. *Annals of the American Association of Geographers*.

Wu, B., Rongrong, Li., & Huang, Bo. (2014) A geographically and temporally weighted autoregressive model with application to housing prices. *International Journal of Geographical Information Science 28(5), 1186-1204.*

Yu, D. (2010) Exploring spatiotemporally varying regressed relationships: the geographically weighted panel regression analysis. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *38*(2), 134–139.