International Journal of Geographical Information Science, 2013
Vol. 27, No. 3, 511–541, http://dx.doi.org/10.1080/13658816.2012.698014

Taylor & Francis
Taylor & Francis Group

The comap as a diagnostic tool for non-stationary kriging models

Paul Harris[a]*, Chris Brunsdon[b] and Martin Charlton[a]

[a]*National Centre for Geocomputation, National University of Ireland, Maynooth, Co. Kildare, Ireland;* [b]*School of Environmental Sciences, University of Liverpool, Liverpool, UK*

(*Received 14 July 2011; final version received 11 May 2012*)

In this study, we demonstrate a novel use of comaps to explore spatially the performance, specification and parameterisation of a non-stationary geostatistical predictor. The comap allows the spatial investigation of the relationship between two geographically referenced variables via conditional distributions. Rather than investigating bivariate relationships in the study data, we use comaps to investigate bivariate relationships in the key outputs of a spatial predictor. In particular, we calibrate moving window kriging (MWK) models, where a local variogram is found at every target location. This predictor has often proved worthy for processes that are heterogeneous, and most standard (global variogram) kriging algorithms can be adapted in this manner. We show that the use of comaps enables a better understanding of our chosen MWK models, which in turn allows a more informed choice when selecting one MWK specification over another. As case studies, we apply four variants of MWK to two heterogeneous example data sets: (i) freshwater acidification critical load data for Great Britain and (ii) London house price data. As both of these data sets are strewn with local anomalies, three of our chosen models are robust (and novel) extensions of MWK, where at least one of which is shown to perform better than a non-robust counterpart.

Keywords: visualisation; coplot; robust; local variogram; geographically weighted

1. Introduction

Geostatistical methods have been widely used in the analysis of spatial data, both as a tool for prediction and as a means for investigating the spatial structure of measurements made of a geographical process. However, one important aspect of this is that some processes may exhibit heterogeneity in this structure. For this reason, a number of localised approaches have been proposed. However, it is important to ensure that the correct approach is used in any particular situation and a key task is deciding on an appropriate specification of a particular model and the interpretation of its output. Diagnostic tools play an important role in reaching this decision; some are based on statistical models, whereas others are more exploratory. In this study, we focus on the latter, where we demonstrate a novel use of comaps (Brunsdon 2001, Brunsdon *et al*. 2007) to explore the performance, specification and parameterisation of a non-stationary predictor from a spatial viewpoint. We show that our understanding of this (complex) predictor whose parameters are allowed to vary across space can benefit strongly from the use of comaps, as a diagnostic tool, enabling a more informed choice when selecting one model specification over another.

*Corresponding author. Email: paul.harris@nuim.ie

The comap is a direct geographical variant of the coplot (Cleveland 1993) and allows the spatial investigation of the relationship between two geographically referenced variables via conditional distributions. In this study, we do not use comaps to investigate bivariate relationships in the raw data but instead use them to investigate bivariate relationships in the outputs of a spatial predictor. Here we calibrate moving window kriging (MWK) models (Haas 1990, 1996), where a local variogram is found at every target location. These models are designed for heterogeneous processes and most global variogram-based kriging algorithms can be adapted in this manner. This non-stationary adaptation not only has the potential to provide more accurate kriging predictions but also improve their associated estimates of prediction confidence (i.e. provide more realistic kriging variances).

We calibrate four variants of MWK, each with two case study data sets. Three of the MWK models are robust in form, where (i) skewed sample data are Box–Cox transformed, (ii) variograms are estimated robustly and (iii) the prediction data are winsorised. These novel extensions of MWK show promise for our study data sets, both of which are heterogeneous and exhibit local anomalies. These data sets are (a) a freshwater acidification critical load data set for Great Britain (CLAG Freshwaters 1995) and (b) a London house price data set for 1998 provided by the Nationwide Building Society (Fotheringham *et al*. 2002). Both spatial processes are known to suit a local variogram modelling approach (for the critical load data, see Harris *et al*. (2010), and for house price data, see Case *et al*. (2004) and Páez *et al*. (2008)).

Our study is structured as follows. First, we introduce comaps together with spatial kernel density estimation (KDE) (Diggle 1990), the latter being one option for spatial data display in a comap. We also define a selection of geographically weighted (GW) summary statistics (Brunsdon *et al*. 2002) at this juncture, as these are used in various guises throughout this study. We then describe the MWK models, together with how we choose to assess them. Next, we describe and model each case study data set, demonstrating the value of comaps in visualising MWK outputs.[1] Finally, we summarise this research and discuss further work.

2. Comaps and other visualisation techniques

We not only demonstrate the use of comaps for visualising paired outputs from MWK models but also demonstrate some univariate visualisations for context. For the latter, GW summary statistic surfaces can be found for MWK outputs that vary continuously across space (e.g. a variogram model parameter) and spatial KDE surfaces can be found for MWK outputs akin to spatial point processes (e.g. a variogram model type). Common to all techniques is a local theme, requiring the specification of some (isotropic) kernel weighting function. Definitions for all kernels used in this study can be found in Wand and Jones (1995).

2.1. Basic and robust GW summary statistics

For GW summary statistics, a locally defined statistic is determined at grid locations $\mathbf{x}$, so that a surface results, enabling a visual inspection of a certain spatial heterogeneity. GW statistics can not only be calculated for MWK outputs but also enable a preliminary investigation into the heterogeneous nature of the study data itself. Furthermore, the value of applying a robust MWK model can be gauged if both basic and robust GW statistics are calculated and compared. In this respect, we define (moment-based) GW means

and standard deviations (and coefficients of variation) together with their respective robust (quantile-based) counterparts, in GW medians and interquartile ranges (IQRs).

A GW mean is defined as $m(\mathbf{x}) = \sum_{i=1}^{n} w_i z(\mathbf{x}_i) \Big/ \sum_{i=1}^{n} w_i$, a GW standard deviation $s(\mathbf{x}) = \sqrt{\sum_{i=1}^{n} w_i (z(\mathbf{x}_i) - m(\mathbf{x}))^2 \Big/ \sum_{i=1}^{n} w_i}$ and a GW coefficient of variation (CoV) $\mathrm{CoV}(\mathbf{x}) = s(\mathbf{x}) \big/ m(\mathbf{x})$, where $z(\mathbf{x}_1), \ldots, z(\mathbf{x}_n)$ at locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ refers to either the MWK output data or the sample data. For GW medians and IQRs, 25%, 50% and 75% local quantiles are needed. Here a local quantile can be defined as $q(\mathbf{x}) = z(\mathbf{x}_J) + (z(\mathbf{x}_{J+1}) - z(\mathbf{x}_J)) \frac{p(\mathbf{x}) - w_J^*}{w_{J+1}^* - w_J^*}$; where J is the index of the largest $z(\mathbf{x}_i)$, not exceeding $q(\mathbf{x})$; $w_1 + \cdots + w_j = w_J^*$ is the probability that $z(\mathbf{x}) \leq q(\mathbf{x})$ when $q(\mathbf{x}) = z(\mathbf{x}_J)$; $w_1 + \cdots + w_{J+1} = w_{J+1}^*$ is the probability that $z(\mathbf{x}) \leq q(\mathbf{x})$ when $q(\mathbf{x}) = z(\mathbf{x}_{J+1})$ and $p(\mathbf{x})$ is the probability that $z(\mathbf{x}) < q(\mathbf{x})$. For this study, let the weights w_i accord to a bisquare kernel with a user-specified adaptive bandwidth.

2.2. Spatial KDE

Spatial KDE allows the identification of clusters where a particular MWK model output is most intense. The KDE of $f(\mathbf{x})$ at a location $\mathbf{x}$ can be found using:

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^2} \sum_{i=1}^{n} k\left(\frac{|\mathbf{x} - \mathbf{x}_i|}{h}\right) \tag{1}$$

where $k(\cdot)$ is some spatially defined kernel function, h is the bandwidth that controls the smoothing and the model output data are located at $\mathbf{x}_1, \ldots, \mathbf{x}_n$. We specify our study KDE surfaces (outside of those used within a comap) using a quartic kernel with a user-specified fixed bandwidth. For KDE details, see Wand and Jones (1995).

2.3. Comaps

For bivariate visualisations of MWK outputs, comaps can be constructed.[2] The comap is a geographical variant of the coplot for assessing conditional distributions, where the underlying concept is based on the idea of using *small multiples* to emphasise variation in pattern (Tufte 1990). In the coplot, the relationship between a pair of variables is plotted conditional to either (i) the third variable or (ii) the third and fourth variables together. Thus, any relationship seen in the initial pair of variables is assessed for dependence on the value(s) of the conditioning variable(s). Each set of variables is a subset of the whole data set, selected so that each conditioning variable lies within a given range. Principles of the subsetting are governed by how the range of each subset overlaps with each adjoining subset and that each subset should contain the same number of data points. These principles are applied in order to ensure patterns are visible (overlapping ensures a reasonable number of data points are used in each plot) and that no plot is more prone to outliers or unusual patterns than any other due to sample size (i.e. by making each plot contain the same number of points). These steps provide an intuitive and reliable tool for data exploration. Each plot window in the coplot also represents the same extent, thus allowing an effective comparison between them.

In the spatial setting, the coplot is specified in the same manner, except that the coordinate data are the first pair of variables conditioned on a single variable or a pair of variables together. A coplot of this form can be extended to a comap by simply adding a map outline to provide context to each plot window of the subsetted coordinate data. A refinement of this basic comap is to then use spatial KDE to visualise the intensity of the same geographic locations. Here the KDE is specified using a normal density kernel function with a bandwidth selected automatically. This automated procedure is designed to select a conservative bandwidth that is likely to produce a degree of over-smoothing (see discussions given in Brunsdon (2001)).

Thus, for the coplot, the output is a panel (or matrix) of plots, whereas for the comap, it is a panel of maps, where the number of plots or maps is user-specified. As an example, we present a spatial coplot and two comaps in Figure 1 (each specified with a 3×3 panel of plot or map windows), using data from our first case study (see Section 5). Observe that the coordinate data are conditioned on a *pair of variables of interest*; and this will be the case in all our comaps, as we are primarily interested in bivariate visualisations. We visualise the relationship between freshwater acidification critical load data and a covariate (geological sensitivity) at 189 sites across Great Britain. A critical load reflects a freshwater site's capacity to buffer the input of strong acid anions. For sites where deposition exceeds the critical load, harmful levels of acidification are expected. Geological sensitivity reflects the nature of a site's catchment geology in its ability to buffer against acidification, where low values have a low buffering capacity and vice-versa (Harris and Juggins 2011).

Thus, for each coplot or comap of Figure 1, the first column relates to low critical loads (0–4 keq H^+ $ha^{-1}year^{-1}$); the second column to critical loads between 1.5 and 9.2 keq H^+ $ha^{-1}year^{-1}$ and the third column to critical loads between 4 and 26.4 keq H^+ $ha^{-1}year^{-1}$. Similarly, the first row relates to catchments with a *high* geological sensitivity (2–4 units), the second row to catchments with an *average* geological sensitivity (1–3 units) and the third row to catchments with a *low* geological sensitivity (1–2 units). Clearly, the spatial coplot (Figure 1a) is little different to the basic comap (Figure 1b), the key difference being context in that the subsetted coordinate data are set within an outline of Great Britain. Finding a spatial pattern for critical load and geological sensitivity is difficult with these visualisations. However, a certain clarity emerges in the second comap (Figure 1c), where KDE surfaces are specified. Here the bottom-left map window depicts a relatively intense clustering, suggesting that critical loads in north-west (NW) Scotland tend to be low valued together with catchments that are geologically sensitive to acidification. Conversely, the KDE surface in the top-right map window, which suggests the reverse is true in Central England. Thus, depending on the nature of the associated deposition rates, freshwaters in NW Scotland are likely to warrant some form of preferential management.

3. Non-stationary kriging models

3.1. Moving window kriging

For MWK, a local variogram is determined at every target location in this fully automatic, continuous and locally adaptive geostatistical technique. At each location, the variogram parameters are used to calibrate a kriging algorithm to provide a prediction and an estimate of its error variance (the kriging variance). MWK models can extend most standard (stationary variogram) algorithms; for example Haas (1990, 1996) constructs lognormal, regression and co-kriging versions; Lloyd and Atkinson (2002) and Pardo-Igùzquiza *et al*. (2005) present universal kriging versions and Cattle *et al*. (2002) present

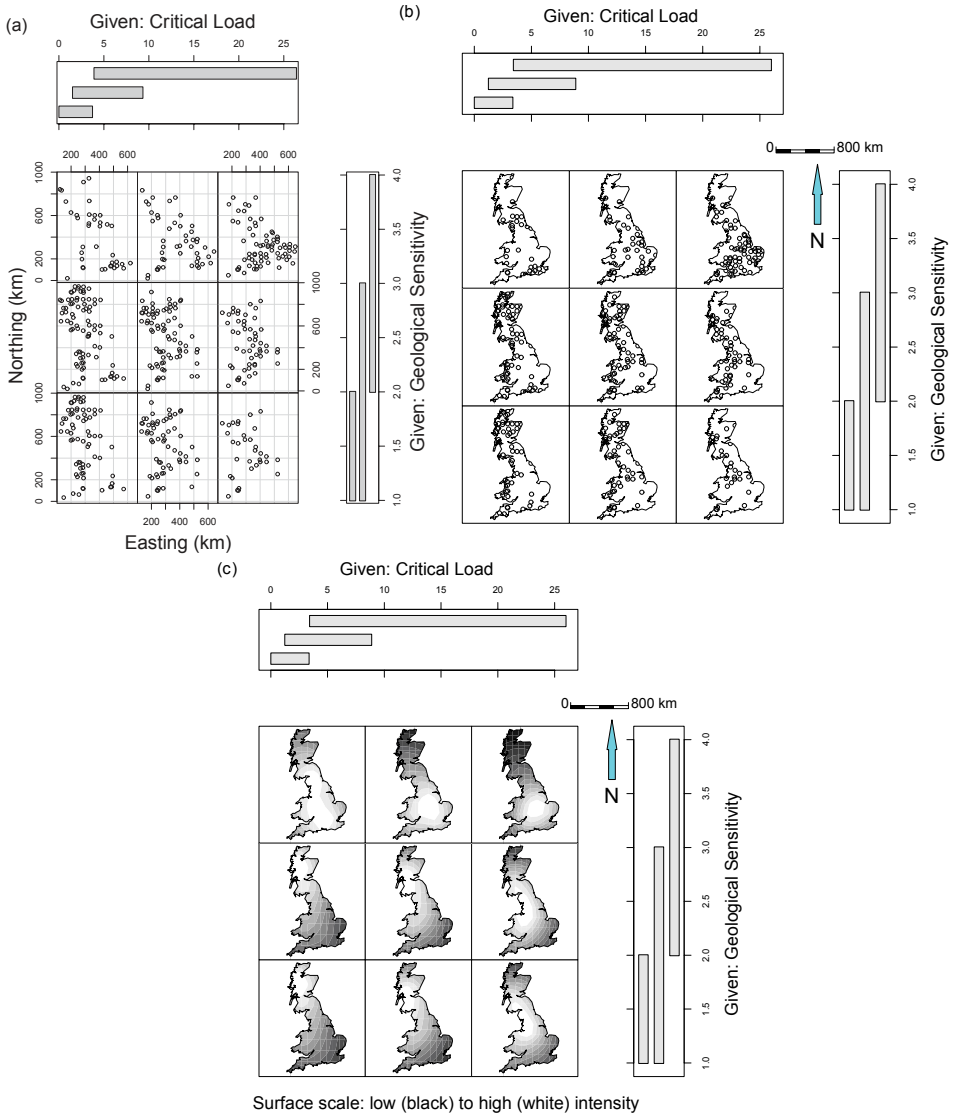

Figure 1. From coplot to comap using critical load and geological sensitivity data: (a) spatial coplot; (b) basic comap with locations (i.e. essentially a spatial coplot with map outlines) and (c) comap with KDE surfaces.

an indicator kriging version. Typically, rudimentary MWK models *globally specify* (i.e. the model specification is fixed in all locations): for example, an algorithm could use a classic variogram estimator, a weighted least squares (WLS) variogram model fit and an exponential variogram model everywhere, with only the numerical parameters varying. Alternative *globally specified* components might include maximum likelihood parameter estimation (Pardo-Igùzquiza *et al*. 2005, Lloyd 2010). Furthermore, models can *locally specify* (i.e. the algorithm or model specification varies across space), so that the variogram estimator, the variogram model-type, the type of data detrending (Haas 1990, 1996,

Pardo-Igùzquiza *et al*. 2005) or the type of data transform (Haas 2002) may change locally, rather than just the numerical parameters.

3.2. Robust MWK

For this study, we calibrate a new set of robust MWK models. Study models are (i) a basic MWK model extending simple kriging (SK; Schabenberger and Gotway 2005, p. 223–225), where all inputs (except for the variogram model-type) are *globally specified* (Mod-1); (ii) a robust moving window SK model, where a Box–Cox transform, a robust variogram and winsorised prediction data are *globally specified* (Mod-2); (iii) a robust moving window SK model, where the Box–Cox transform of (ii) is *locally specified* (Mod-3) and (iv) a robust moving window SK model, where *all* robust specifications of (ii) are *locally specified* (Mod-4). Observe that SK is chosen as the basic kriging algorithm as it enables simpler back-transform corrections than ordinary kriging (OK). Unlike OK, SK assumes a known mean, which we estimate using the sample data.[3]

Our robust MWK models are entirely novel, adapting the robust (stationary variogram) algorithms of Hawkins and Cressie (1984), Costa (2003) and Marchant *et al*. (2010). As an MWK calibration only uses local data subsets, it is likely that at some locations, its local variography, its locally weighted predictions and its kriging variances are more (adversely) influenced by data non-normality and outliers than a global counterpart. In this respect, the application of a robust MWK model may be of value for processes that are heterogeneous with local anomalies. Both of our case study data sets exhibit these properties. However, as is often the case, it is not known beforehand which particular robust MWK specification will suit our study data sets? Furthermore, standard model performance diagnostics (Section 4.1) are sometimes inconclusive (Sections 5.1 and 6.1) suggesting models perform more or less equally. It is precisely in these cases, that we aim to demonstrate the value in exploring MWK outputs using comaps, so that key insights into a model's spatial behaviour can be gained. Comap results can supplement those found using standard diagnostics, enabling an informed choice when selecting an MWK specification.

3.3. The four study models

3.3.1. Basic MWK: Mod-1

In this model, SK is applied within a window (or neighbourhood) using variogram parameters found at the same spatial scale. Window size is specified in an adaptive form, with each window extending to the nearest N model calibration observations (and is given as a percentage of the full model calibration data set). A two-stage variographic procedure is adopted, where a classic (isotropic) variogram is first estimated and then modelled to get a smooth representation of spatial dependence. The classic variogram estimator can be defined as:

$$2\,\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(i,j)\,\in N(h)} \left\{ z(\mathbf{x}_i) - z\left(\mathbf{x}_j\right) \right\}^2, \quad \text{for} \quad h > 0 \tag{2}$$

where $z(\mathbf{x}_1), \ldots, z(\mathbf{x}_n)$ denote observations from a constant mean spatial process at locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and $N(h)$ is a count of pairs of indices such that $(i,j) \in N(h)$ if $\left| \left\| \mathbf{x}_i - \mathbf{x}_j \right\| - h \right| < \varepsilon$ (where this approximation reflects some user-specified tolerance ε

on the lag distance h). To interpret an estimated variogram, nugget and Matérn (1960) variogram model types are specified, which can be defined as:

$$\gamma_{\mathrm{NUG}}(h) = c_0 + c_1 \qquad \text{and} \tag{3}$$

$$\gamma_{\mathrm{MAT}}(h) = c_0 + c_1 \left(1 - \frac{1}{2^{v-1}\Gamma(v)} \left(\frac{h}{a}\right)^{v} K_v \left(\frac{h}{a}\right)\right), \quad \text{for} \quad h > 0 \tag{4}$$

where $\gamma(h) = 0$ for $h = 0$. Here a is a correlation range parameter, v is a smoothing parameter, K_v is a modified Bessel function, Γ is the gamma function and c_0 and c_1 are partial sills that reflect small- and large-scale variations, respectively. The Matérn function is very flexible, where the higher is the value of v, the smoother is the process. For this study, v is fixed beforehand at 0.5, 1.0 or 1.5. This effectively provides four (not two) variogram models to choose from in total.

Thus, at each target location, all four variogram models are fitted automatically to a local estimated variogram, using the WLS approach of Zhang *et al*. (1995). Here a series of sensible heuristics are used to define (i) the lag distance intervals, (ii) the minimum number of pairs allowed at the first lag and (iii) the truncation of the variogram estimator at long (unreliable) lag distances; together with (iv) the variogram model starting parameters for the WLS fit. Such specifications are chosen so as to minimise the occurrence of a poor (or failed) WLS fit and are crucial in determining a good kriging model performance (Lahiri *et al*. 2002). One variogram model is then chosen that provides the smallest weighted sum of squares (i.e. the variogram model is *locally specified*) and the corresponding parameters are used to solve an SK system of equations. Similar automated procedures are used in many other MWK studies (Haas 1990, 1996).

3.3.2. Robust MWK: Mod-2

In our first robust model, Mod-1 is (sequentially) adapted as follows: (i) *globally* Box–Cox transform the data; (ii) *locally* estimate and model (isotropic) robust variograms using the transformed data; (iii) *locally* apply (the four stages of) a robust form of SK, where the robust variogram model parameters are used to find Winsorised data sets; (iv) back-transform the robust SK results to the original data space. Observe that all robust specifications are *globally-defined.* The techniques to conduct these adaptations are as follows:

The Box–Cox method (Box and Cox 1964) transforms the data to univariate normality according to:

$$z_t = \begin{cases} \frac{z^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln(z) & \lambda = 0 \end{cases} \tag{5}$$

where z_t is the transformed value and z is the value to be transformed. We estimate the parameter λ using maximum likelihood based on an assumption that the transformed data are Gaussian and correlated.[4]

Variography and kriging with transformed data is no different to that for raw data. However, to account for a bias when back-transforming, the following *corrected* back-transforms are used (Kitanidis and Shen 1996). When $\hat{\lambda} \neq 0$, an

unbiased back-transform for a prediction at any location $\mathbf{x}$ is taken as $\hat{z}_{\mathrm{BCK}}(\mathbf{x}) = \beta^{\alpha}\left[\beta^2 + \alpha\,\sigma^2_{\mathrm{BCKT}}(\mathbf{x})\right]$, where $\beta = 1 + \hat{\lambda}\,\hat{z}_{\mathrm{BCKT}}(\mathbf{x})$, $\alpha = \left(1\Big/\hat{\lambda}\right) - 2$, $\hat{z}_{\mathrm{BCKT}}(\mathbf{x})$ is the SK prediction in transformed-space and $\sigma^2_{\mathrm{BCKT}}(\mathbf{x})$ is the SK variance in transformed space. Alternatively when $\hat{\lambda} = 0$, an unbiased back-transform results in $\hat{z}_{\mathrm{BCK}}(\mathbf{x}) = \exp\left[\hat{z}_{\mathrm{BCKT}}(\mathbf{x}) + \sigma^2_{\mathrm{BCKT}}(\mathbf{x})\big/2\right]$. An unbiased back-transform for the corresponding kriging variance is as follows:

$$\sigma^2_{\mathrm{BCK}}(\mathbf{x}) = \sigma^2_{\mathrm{BCKT}}(\mathbf{x})\ \left[\left(\sigma^2_{\mathrm{BCKT}}(\mathbf{x})\big/8\right) + \beta^2\right],\ \text{if}\ \hat{\lambda} = 0.5 \tag{6}$$

$$\sigma^2_{\mathrm{BCK}}(\mathbf{x}) = \exp\left[\left(2\,\hat{z}_{\mathrm{BCKT}}(\mathbf{x}) + \sigma^2_{\mathrm{BCKT}}(\mathbf{x})\right)\right] \times \left[\exp\left(\sigma^2_{\mathrm{BCKT}}(\mathbf{x})\right) - 1\right],\ \text{if}\ \hat{\lambda} = 0 \tag{7}$$

$$\text{and}\ \sigma^2_{\mathrm{BCK}}(\mathbf{x}) = \beta^{\delta}\,\sigma^2_{\mathrm{BCKT}}(\mathbf{x})\ \text{for all other values of}\ \hat{\lambda},\ \text{where}\ \delta = \left(2\Big/\hat{\lambda}\right) - 2 \tag{8}$$

These analytical back-transforms are sensitive to values of the kriging variance, and, as a consequence, poor or unusual kriging outputs can often be attributed to this.

To guard against the effects of outliers on variogram estimation, we use the robust estimator proposed by Cressie and Hawkins (1980), although alternatives could be used (Lark 2000). This robust estimator can be defined as:

$$\hat{\gamma}_r(h) = \frac{\left[\frac{1}{|N(h)|}\sum_{(i,j)\in N(h)}\left|z(\mathbf{x}_i) - z\left(\mathbf{x}_j\right)\right|^{\frac{1}{2}}\right]^4}{\left(0.914 + \left(\frac{0.988}{N(h)}\right) + \left(\frac{0.045}{N^2(h)}\right)\right)}, \quad \text{for} \quad h > 0 \tag{9}$$

where $\hat{\gamma}_r(h)$ is an approximately unbiased estimator of $\hat{\gamma}(h)$.

To complement this use of robust variography, we can similarly reduce the influence of outlying observations in the prediction part of the SK algorithm by a process of *Winsorisation* (Tukey 1962). Thus, generally following Hawkins and Cressie (1984), the four stages of this robust kriging algorithm at a location $\mathbf{x}$ are as follows:

A. Use a robust variogram to find the usual (mean-based) SK prediction $\hat{z}(\mathbf{x}) = \sum_{i=1}^{N}\eta_i(\mathbf{x})\,z(\mathbf{x}_i)$ and its standard error $\sigma(\mathbf{x})$ (where for MWK, $n > N$).
B. Find a corresponding (robust and weighted) median-based SK prediction $\tilde{z}(\mathbf{x})$ using the same SK weights $\eta_i(\mathbf{x})$ as used in (A).
C. Winsorise the sample data used in the SK predictions of (A) and (B) by replacing $z(\mathbf{x}_i)$ with:

$$z_w(\mathbf{x}_i) = \begin{cases} \tilde{z}(\mathbf{x}) + b\sigma(\mathbf{x}) & \text{if} \quad z(\mathbf{x}_i) - \tilde{z}(\mathbf{x}) > b\sigma(\mathbf{x}) \\ z(\mathbf{x}_i) & \text{if} \quad |z(\mathbf{x}_i) - \tilde{z}(\mathbf{x})| \le b\sigma(\mathbf{x}) \\ \tilde{z}(\mathbf{x}) - b\sigma(\mathbf{x}) & \text{if} \quad z(\mathbf{x}_i) - \tilde{z}(\mathbf{x}) < -b\sigma(\mathbf{x}) \end{cases} \tag{10}$$

where b is a constant that controls the degree of data editing. The higher the value of b, the fewer the number of changes to the original $z(\mathbf{x}_i)$ data, that is, $z_w(\mathbf{x}_i)$ tends to $z(\mathbf{x}_i)$. It is recommended that $1.5 < b < 2.5$ and we (conservatively) choose $b = 2.5$. Further research could investigate this choice of b more closely (Costa 2003, Marchant *et al*. 2010) where it could be *locally specified.*

D. Use a classic variogram to find a mean-based SK prediction $\hat{z}_w(\mathbf{x}) = \sum_{i=1}^{N} \varphi_i(\mathbf{x})\, z_w(\mathbf{x}_i)$ and its standard error $\sigma_w(\mathbf{x})$, where $z_w(\mathbf{x}_i)$ has replaced $z(\mathbf{x}_i)$ in the variography and the SK prediction ($\varphi_i(\mathbf{x})$ are the new SK weights). These are the robust SK results.

Observe that at each target location four variogram models are fitted automatically to a local robust variogram estimator (using pre-winsorised data) and four variogram models are fitted automatically to a local classic variogram estimator (using winsorised data). These operations, together with the use of transformed data, ensure a considerable increase in flexibility over the basic MWK model (Mod-1).

3.3.3. *Robust MWK: Mod-3*

Our next robust model is the same as Mod-2, except that the Box–Cox transform is now *locally specified.* Here we modify the Box–Cox procedure so that the local calibration data subset relating to location $\mathbf{x}$ is transformed to normality only if $\hat{\lambda}(\mathbf{x}) < 0.55$. This pragmatic approach enables robust SK with raw data to be used at locations where outliers create only a mild skew in the local distribution, which in turn should reduce the opportunities for an unrealistic back-transform to occur. Furthermore if $\hat{\lambda}(\mathbf{x}) < 0.05$, $\hat{\lambda}(\mathbf{x})$ is taken to equal 0 and a log transform is applied; if $0.05 \leq \hat{\lambda}(\mathbf{x}) < 0.45$, the transform is applied as is; and if $0.45 \leq \hat{\lambda}(\mathbf{x}) < 0.55$, $\hat{\lambda}(\mathbf{x})$ is taken to equal 0.5 and a square root transform is applied. Log and square root transforms are promoted as they have a more natural interpretation than a closely related Box–Cox transform.

3.3.4. *Robust MWK: Mod-4 (hybrid)*

Our third robust model is again similar to our first one, except that *all* robust specifications are *locally defined.* This model is effectively a hybrid; where at some locations Mod-1 is applied, whereas at others, Mod-3 is applied. The rationale for this model is that non-robust SK predictions should be preferred at locations where local calibration data subsets conform well to a normal distribution model without local anomalies. At such locations, robust SK predictions should be avoided as (i) back-transforms can produce unusual results, (ii) our chosen robust variogram estimator is biased and less efficient than the classic estimator (Schabenberger and Gotway 2005, p. 160–161) and (iii) data winsorisation adds an unnecessary layer of complexity.

Thus, for this robust model, we simply need to calculate a local diagnostic at a target location $\mathbf{x}$, which can be used to indicate the likely presence of outliers (and non-normality) in the corresponding local calibration data subset. This in turn allows a non-robust or robust SK prediction to be chosen. There are various options for this diagnostic, but for simplicity we use adjusted boxplot statistics (Hubert and Vandervieren 2008). Here the existence of at least one observation that lies beyond the outer fences of the local boxplot indicates a robust SK prediction. An adjusted boxplot is chosen as it uses a robust measure of skewness to determine its whiskers (and so lessens the chance of a false positive identification). Observe that this procedure depends on how we specify the outer fences of the boxplot. Here we use defaults, but further research could vary this choice (Reimann *et al*. 2005).

4. Model assessment

4.1. Standard diagnostics for model performance

For actual $z(\mathbf{x}_v)$ and predicted $\hat{z}(\mathbf{x}_v)$ data, a model's prediction accuracy is measured by:

$$\text{MPE} = \left(1/M\right) \sum_{v=1}^{M} \left\{ z\left(\mathbf{x}_v\right) - \hat{z}\left(\mathbf{x}_v\right) \right\} \tag{11}$$

$$\text{RMSPE} = \sqrt{\left(1/M\right) \sum_{v=1}^{M} \left\{ z\left(\mathbf{x}_v\right) - \hat{z}\left(\mathbf{x}_v\right) \right\}^2} \tag{12}$$

$$\text{MAPE} = \left(1/M\right) \sum_{v=1}^{M} \left| z\left(\mathbf{x}_v\right) - \hat{z}\left(\mathbf{x}_v\right) \right| \tag{13}$$

where expressions (11)–(13) are the mean prediction error (MPE), the root mean squared prediction error (RMSPE) and the mean absolute prediction error (MAPE), respectively, and M is the size of the model validation data set. A model's (overall) prediction uncertainty accuracy is measured using the mean squared deviation ratio (MSDR):

$$\text{MSDR} = \left(1/M\right) \sum_{v=1}^{M} \left(\left\{ z\left(\mathbf{x}_v\right) - \hat{z}\left(\mathbf{x}_v\right) \right\}^2 \Big/ \sigma^2\left(\mathbf{x}_v\right) \right) \tag{14}$$

An MSDR value <1 implies that the kriging variances $\sigma^2(\mathbf{x}_v)$ tend to over-estimate the squared prediction errors (and vice versa).[5] Similarly, a more locally orientated diagnostic is taken as the linear correlation coefficient between the absolute prediction errors $\left| z(\mathbf{x}_v) - \hat{z}(\mathbf{x}_v) \right|$ and the kriging standard errors $\sigma\left(\mathbf{x}_v\right)$. This correlation (ERR-CORR) should be positive and moderately strong if the values of $\sigma\left(\mathbf{x}_v\right)$ actually reflect the local variability that is present in the sample data.[6]

Prediction confidence interval (PCI) accuracy can be assessed using coverage probabilities (Goovaerts 2001). For example, if symmetric 95% PCIs were calculated at each validation site (i.e. $\hat{z}(\mathbf{x}_v) \pm 1.96\,\sigma\left(\mathbf{x}_v\right)$ under the Gaussian assumption), then a correct modelling of local uncertainty would entail that there is a 0.95 (expected coverage) probability that the actual value $z\left(\mathbf{x}_v\right)$ falls within the interval. If a coverage probability is found for a range of symmetric PCIs (say from a 1% to a 99% PCI in increments of 1%) and the results plotted against the probability interval p, then an accuracy plot is found which should follow the 45°, $x = y$ line. In this study, we do not present accuracy plots but instead summarise them via the G-statistic, defined as:

$$\text{G-STAT} = 1 - \int_0^1 \left[3a\left(p\right) - 2\right]\left[\bar{\xi}\left(p\right) - p\right] \mathrm{d}p \tag{15}$$

where $\bar{\xi}$ is the fraction of actual values falling in the PCI, and a value of 1 is sought. The indicator function $a(p)$ is defined as $a(p) = \begin{cases} 1 & \text{if } \bar{\xi}(p) \geq p \\ 0 & \text{otherwise} \end{cases}$, which implies that twice the importance is given to deviations when $\bar{\xi}\left(p\right) < p$. For cases where two models provide similar accuracy plots or G-STAT values, one model can be preferred if its PCI widths containing the actual value are narrower (i.e. more precise). Here the corresponding PCI

width plots can be constructed and compared. Again we do not present PCI width plots, but to act as a rough summary of this plot, a mean PCI width (M-PCI-W) for all p is reported, which should be as small as possible. Thus, a model's G-STAT and M-PCI-W values should always be viewed in conjunction, as a strong G-STAT value is of little use if it is coupled with a poor M-PCI-W value (and vice versa). Observe that the given prediction uncertainty accuracy diagnostics must be viewed in the context that the kriging standard error $\sigma(\mathbf{x}_v)$ is a statistical concept defined as the standard deviation of the error at a location $\mathbf{x}_v$, and its use requires the assumption of multivariate normality (at the scale of the kriging window). That is, we adopt a classical geostatistics viewpoint in this respect (David 1988).

4.2. *Model performance: window size and naive models*

Crucial to MWK is the spatial scale at which the local variography takes place. If the window is too small, then the variography tends to be unreliable as information is limited. If the window is too large, the output of MWK will tend to that of its standard kriging counterpart, and thus offers no benefit if the underlying process is not stationary. Unfortunately, there is no easy way to choose a window size that can be considered optimal (Haas 1996, 2002). Commonly, a model's window size that is optimal with respect to prediction accuracy (i.e. via MPE, RMSPE and MAPE values) does not coincide with a model's window size that is optimal with respect to prediction uncertainty accuracy (i.e. via MSDR, ERR-CORR, G-STAT and M-PCI-W values). Haas (2002) tries to address this problem by using a multi-criteria objective function, where a range of model performance diagnostics (similar to those used in this study) are weighted according to the user's relative priorities and then summed to give a single diagnostic. The obvious drawback to this approach is deciding on the weights. Given these problems, we present the performance results of our MWK models using a range of window sizes. From this, we select a few MWK models for further (visual) scrutiny; ones that are *judged* to offer a good compromise between prediction accuracy and prediction uncertainty accuracy.

We also present the performance results for four naive kriging models (for Mod-1 and Mod-2 only), where 100% window sizes are specified for the variography, two of which have different window sizes for the prediction part (i.e. they assume quasi-stationarity, Journel and Huijbregts 1978, p. 33–34). Such stationary variogram kriging models are presented to show the relative value of adopting a non-stationary adaptation in the first place. Observe that the Box–Cox back-transforms defined in Section 3.3 are implicitly designed to cater for a non-stationary variogram, but only if it is assumed to vary proportionally across space (i.e. the global variogram is effectively locally re-scaled) (Chilès and Delfiner 1999, p. 56 and p. 107–108). In this respect, Mod-2 with a 100% window size for its variography could in some cases be expected to perform in a similar manner to a true local variogram kriging model (i.e. MWK where the local variography is more flexibly modelled and thus caters for locally changing variographic shapes/structures). Finally, as a check on the prediction accuracy gain (if any) in using a kriging to a non-kriging method, performance results using a (simple and naive) nearest neighbour predictor are given. Here the 10th nearest neighbour is specified (call this model, NN10).[7]

4.3. *Visualisation categories: performance, specification and parameterisation*

We divide our visualisations into three categories: (i) model performance, (ii) model specification and (iii) model parameterisation. With respect to model performance, the output

from MWK is treated identically to that of a stationary predictor as we are simply visualising the (actual) prediction errors and the kriging standard errors (or estimated prediction errors). We also find novel GW versions of G-STAT and M-PCI-W diagnostics in this respect. For model specification and parameterisation, a wealth of information is possible depending on the complexity of the MWK model. Here we visualise the following outputs with respect to model specification: (i) the number of winsorised data edits found in each local calibration data set (for Mod-3 only); (ii) the number of adjusted boxplot outliers found in each local calibration data set (for Mod-4) and (iii) the location of robust predictions in Mod-4. For model parameterisation, we visualise the following MWK outputs: (A) the variogram parameters $a(\mathbf{x})$ and $c_0(\mathbf{x})$ and (B) the relative structural variability (RSV), where $\mathrm{RSV}(\mathbf{x}) = \left(c_1(\mathbf{x}) \big/ (c_1(\mathbf{x}) + c_0(\mathbf{x}))\right) \times 100\%$ (Schabenberger and Gotway 2005).

5. Case study A: critical loads

Our first case study models a freshwater acidification critical load data set for Great Britain. Here we are modelling (site-specific) threshold data. Resultant critical load predictions (and estimates of uncertainty) would in turn need to be compared with their corresponding deposition (contaminant) prediction model outputs (not modelled here) in order to find critical load exceedances (critical load minus deposition). That is, apply the critical load concept at un-sampled sites (Nilsson and Grennfeld 1988). This dual and interactive prediction methodology should ultimately provide accurate estimates of critical load exceedance risk so that freshwaters can be efficiently managed against harmful levels of acidification to their ecosystems.

This data set consists of 686 values, which split into model calibration and validation subsets of 497 and 189 observations, respectively (Figure 2a). The size and scale of this data set suit non-stationary modelling, where preliminary investigations provided evidence of (i) positively skewed data, (ii) a general trend in critical loads from high to low values in a south-east (SE) to NW direction (Figure 3a and b), (iii) levels of variation (Figures 2a, 3c and d) and skewness changing across space, (iv) spatial outliers, (v) a nested global variogram (Figure 2c) and (vi) structure in the local variography not only varying across space but also varying according to the influence of outlying observations. Figure 3 depicts (a) local averages (using GW means and medians) and (b) local measures of variability (using GW standard deviations and IQRs). Clear regional differences can be observed between the non-robust and robust paired surfaces, which are taken to confirm the presence of a skewed critical load distribution together with some influential, outlying observations.

5.1. Model performance at the validation sites: standard diagnostics

Performance results of our four study models are presented in Table 1, where it is clear that an MWK model does not improve critical load prediction accuracy when compared with a non-robust naive model – Mod-1 with a 100% window size (see MPE, RMSPE and MAPE values). However, this global variogram model performs weakly with respect to prediction uncertainty accuracy (see ERR-CORR, G-STAT and M-PCI-W values) and here a robust MWK model (Mod-3 with a 20% window size) is viewed as the best performer. For prediction uncertainty accuracy via the MSDR diagnostic only, Mod-1 with a 100% window size (for variogram and prediction) performs the best. Mod-3 with a 10% window size is viewed as the best overall predictor, as it offers relatively good prediction accuracy and prediction uncertainty accuracy results.

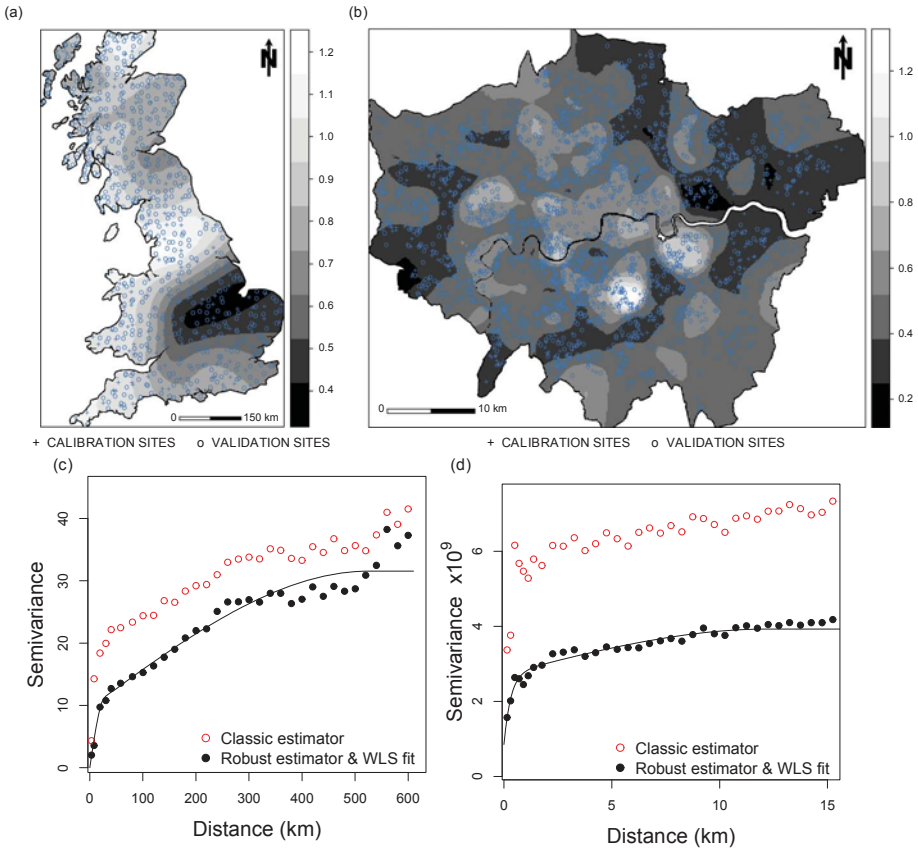

Figure 2. Case study data sets and global variograms for (a and c) freshwater acidification critical load data for Great Britain and (b and d) London house price data, ca. 1998. For context, each data set is shown with a GW coefficient of variation (CoV) surface with bandwidths set at 20% and 2.5% for the critical load and house price data, respectively. Both surfaces use the full case study data sets. Classic and robust variogram estimators shown with a WLS model fit to the robust estimator only.

In summary, there is value in applying MWK to these data, but only marginal promise in using a robust adaptation. This promise may be simply due to statistical noise or variability in the data (which could be verified by using different model calibration and validation subsets and re-fitting the MWK models). We view this as an instance, where an investigation using comaps may be helpful, providing a more local assessment of differences in model performance. As Mod-3 with a 10% window size is considered the best predictor, we choose this model for a visual scrutiny, together with the models Mod-1, Mod-2 and Mod-4 that use the same window size. For context, we first provide examples of univariate visualisations of model outputs using GW statistics and spatial KDE for each of the three visualisation categories. Here we observe various limitations that can be addressed by using comaps instead.

5.2. *Univariate visualisations of MWK outputs*

First, as an example of model performance, GW median surfaces are used to visualise the actual prediction errors and the estimated prediction errors for Mod-1 and Mod-3. GW

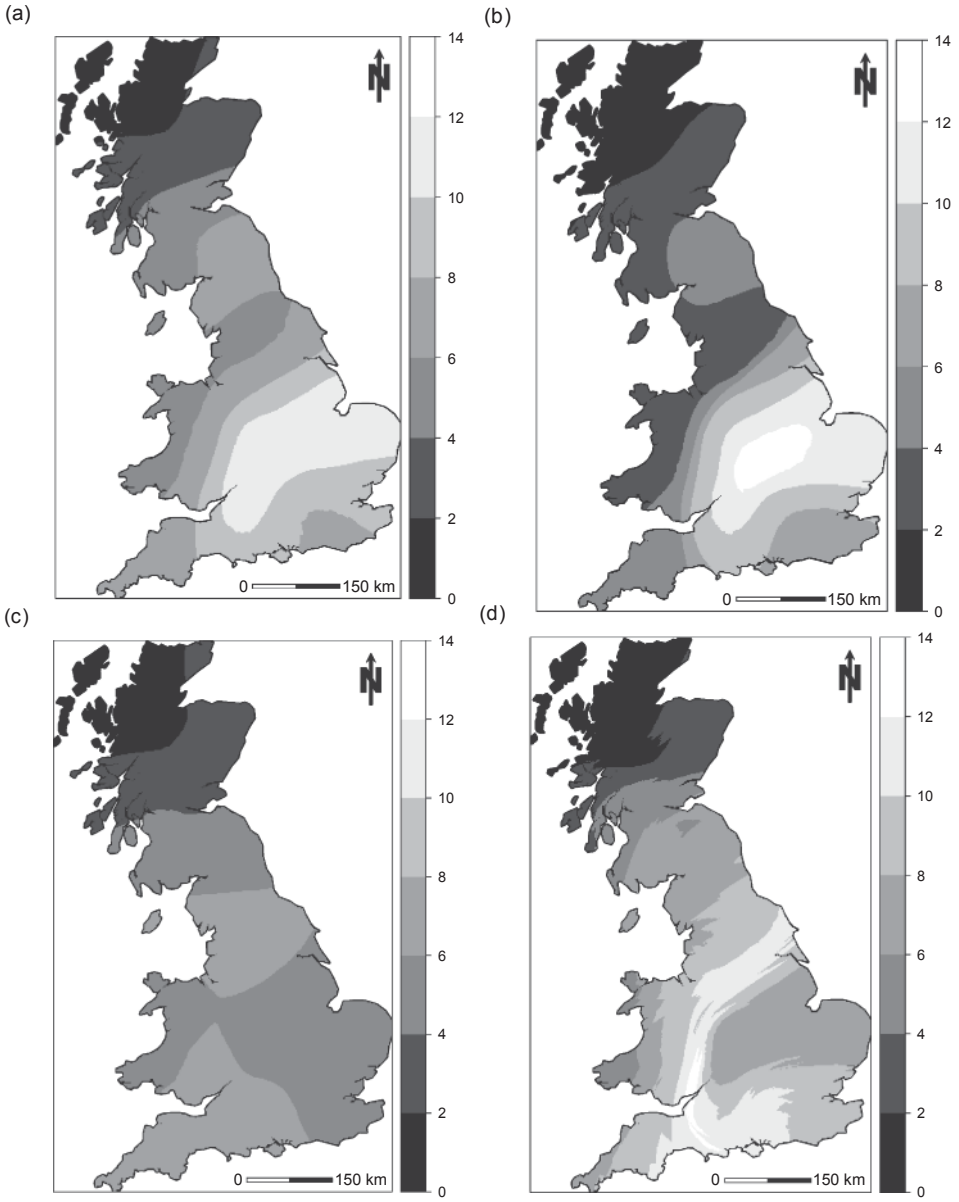

Figure 3. GW summary statistic surfaces for critical load data: (a) mean; (b) median; (c) standard deviation (SD) and (d) interquartile range (IQR). All surfaces use the full case study data sets with bandwidths set at 20%.

medians are specified in order to reduce the effect of a few outlying values. From Figure 4, the largest actual prediction errors for Mod-1 (the non-robust model) are mainly found in two regions: one centred on the coast of NW England and one centred on the coast of Central/Southern England (both areas of known outliers). The robust Mod-3 model has most success in reducing these large prediction errors on the coast of NW England, and this reduction is loosely mirrored in a reduction in the corresponding prediction error

Table 1. Prediction and prediction uncertainty accuracy at the validation sites (critical load data).

Model	Window size: variogram (%)	Window size: prediction (%)	MPE	RMSPE	MAPE	MSDR	ERR-CORR	G-STAT	M-PCI-W
Ideal	–	–	0	→ 0	→ 0	1	→ 1	1	→ 0
NN10	–	–	−0.55	6.59	4.43	2.078	−0.03	0.923	7.28
Mod-1	100[a]	100	−0.11	4.38	3.02	0.973	0.12	0.910	7.01
	100[a]	10	−0.10	4.39	3.02	0.802	0.12	0.887	7.76
	30	30	−0.05	4.41	3.14	0.830	0.35	0.944	7.09
	20	20	−0.12	4.47	3.15	0.816	0.41	0.946	7.04
	10	**10**	**−0.15**	**4.42**	**3.10**	**1.049**	**0.39**	**0.947**	**6.83**
	7	7	−0.06	4.65	3.26	1.116	0.42	0.960	6.68
Mod-2	100[a]	100	0.25	4.50	3.06	2.713	0.30	0.897	4.11
	100[a]	10	0.22	4.49	3.05	2.290	0.31	0.942	4.47
	30	30	0.26	4.49	3.11	1.721	0.36	0.963	5.54
	20	20	0.18	4.52	3.12	1.531	0.39	0.961	5.55
	10	**10**	**0.06**	**4.47**	**3.09**	**1.732**	**0.41**	**0.965**	**5.58**
	7	7	0.11	4.75	3.22	2.359	0.36	0.952	5.47
Mod-3	30	30	0.21	4.50	3.11	1.844	0.35	0.974	5.56
	20	20	0.20	4.51	3.10	1.397	0.41	0.967	5.55
	10	**10**	**0.06**	**4.43**	**3.06**	**1.745**	**0.42**	**0.969**	**5.64**
	7	7	0.20	4.64	3.14	2.172	0.37	0.959	5.68
Mod-4	30	30	−0.02	4.42	3.11	1.141	0.37	0.972	6.76
	20	20	−0.05	4.53	3.16	1.036	0.42	0.975	6.80
	10	**10**	**−0.07**	**4.40**	**3.06**	**1.247**	**0.41**	**0.962**	**6.60**
	7	7	0.01	4.60	3.20	1.431	0.41	0.975	6.50

Notes: Models highlighted in bold are chosen for further scrutiny. Window sizes smaller than 7% suffered from calibration difficulties. [a]These naive models are the equivalent stationary (global) variogram kriging models.

estimates. Clearly, these model outputs are better investigated in pairs, and as such we repeat this study using comaps in Section 5.3.

As an example of model specification, it is important to know where a robust prediction is used in Mod-4 (the hybrid) and where it is not. Thus, we use KDE on the locations of robust predictions, where it appears that they are predominantly specified in Scotland and Central/Eastern England (Figure 5a). This behaviour is, in part, counter-intuitive to the prediction error surfaces given in Figure 4 and may explain why Mod-4 does not perform significantly better than its constituent models, Mod-1 and Mod-3 (see Table 1). Thus, for a better understanding of how we have specified our model in this respect, a comap can be used to investigate the paired relationship between the number of winsorised data edits at each target location for Mod-3 (which can be viewed as spatial outlier detection) and the number of adjusted boxplot outliers at each target location for Mod-4 (which can be viewed as aspatial outlier detection). This comap is also described in Section 5.3.

As an example of model parameterisation, mapping $c_0(\mathbf{x})$ may provide local clues to levels of measurement error and/or unmeasured small-scale variability. We can only map this parameter for models where the variography is conducted in either the original data space or the transformed data space, but not some mixture. As such, only Mod-1 and Mod-2 are investigated via GW mean surfaces (Figure 5b and d). For both models, high values of $c_0(\mathbf{x})$ are generally found in the same areas that result in large prediction errors (i.e. outliers create a heightened local discontinuity in the critical load process). As would be expected, Mod-2 reduces this influence of outliers on the estimates of $c_0(\mathbf{x})$, primarily in

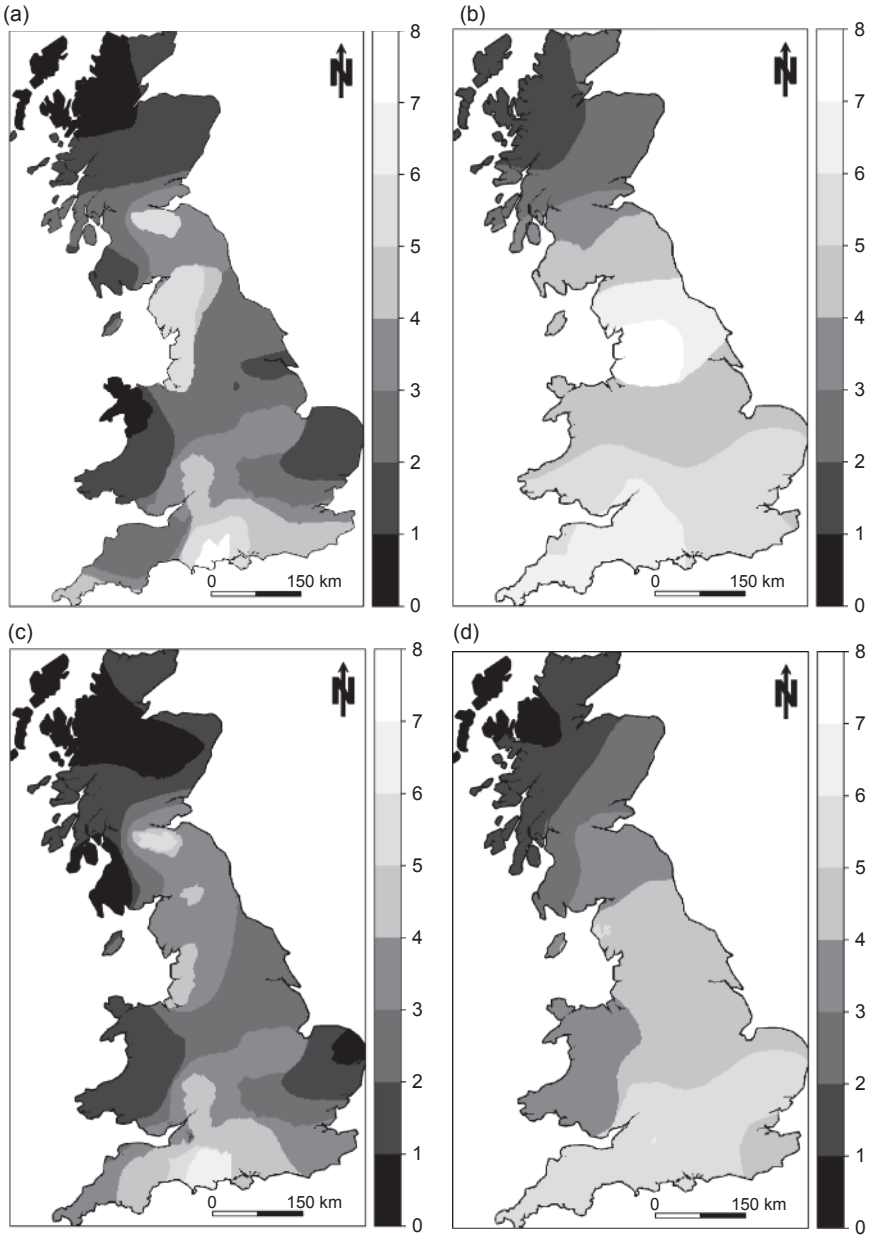

Figure 4. GW median surfaces of the actual (absolute) prediction errors, $\left| z(\mathrm{x}_v) - \hat{z}(\mathrm{x}_v) \right|$ and the estimated prediction standard errors, σ (x_v), for Mod-1, (a) and (b), respectively, and for Mod-3, (c) and (d), respectively. Surfaces are specified with a bandwidth of 10%. All surfaces are for the critical load data.

NW England. Observe that any differences between the pre- and post-winsorised surfaces for Mod-2 relate to a robust MWK model where only robust variograms are specified and winsorisation does not occur. Greater insight into local variographic structure can be found by using comaps to investigate the spatial relationship between $a(\mathbf{x})$ and RSV($\mathbf{x}$) (see Section 5.3).

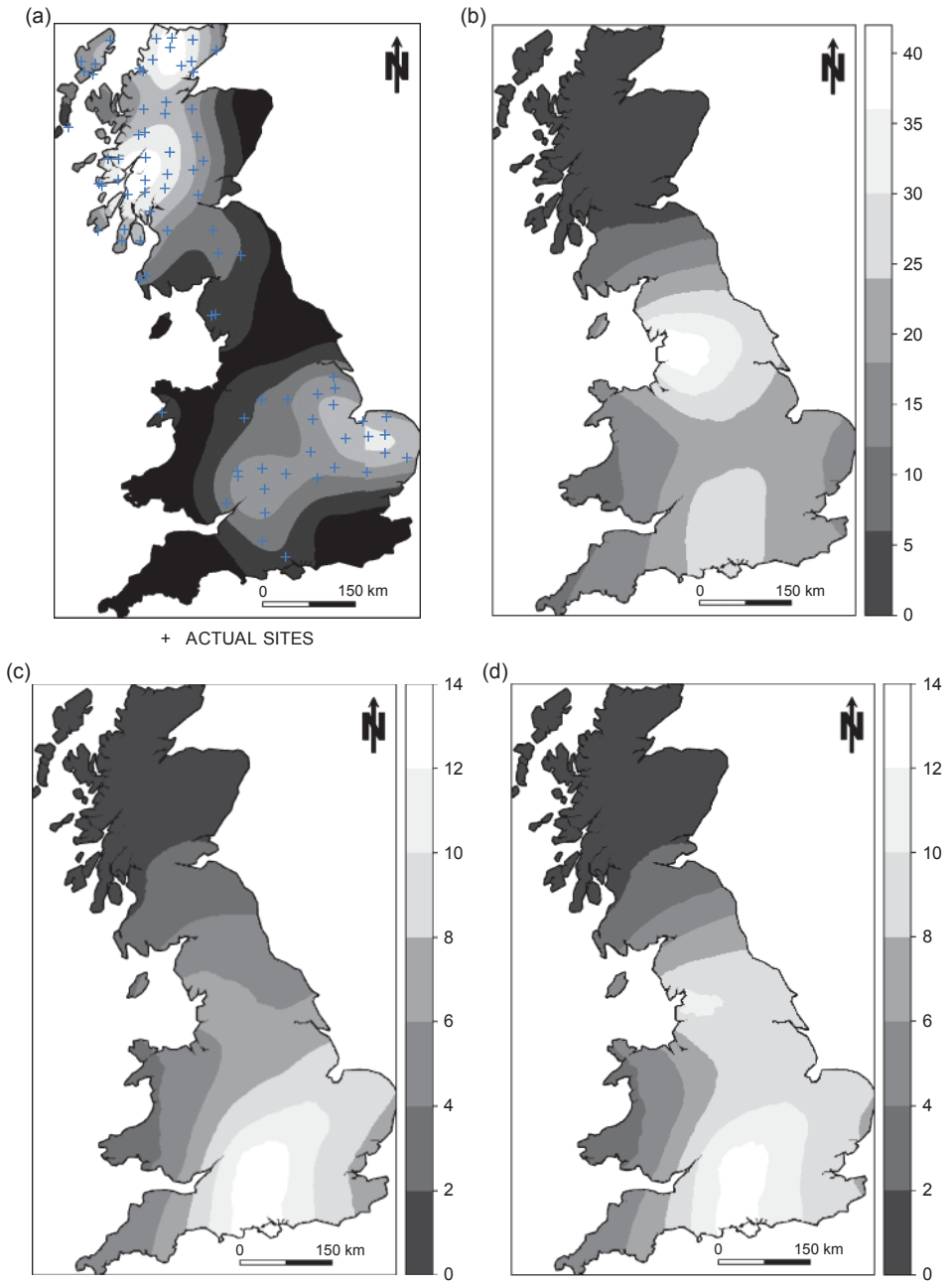

Figure 5. (a) Intensity surface for a robust prediction in Mod-4 (with black to white reflecting low to high intensity, respectively, and KDE specified with a fixed bandwidth of 75 km). GW mean surfaces for the local nugget variances $c_0(x)$ of (b) Mod-1; (c) Mod-2, pre-winsorisation, and (d) Mod-2, post-winsorisation (each specified with an adaptive bandwidth of 20%). Variography for pre-winsorisation uses robust variograms to transformed data. Variography for post-winsorisation uses classic variograms to transformed Winsorised data. All surfaces are for the critical load data.

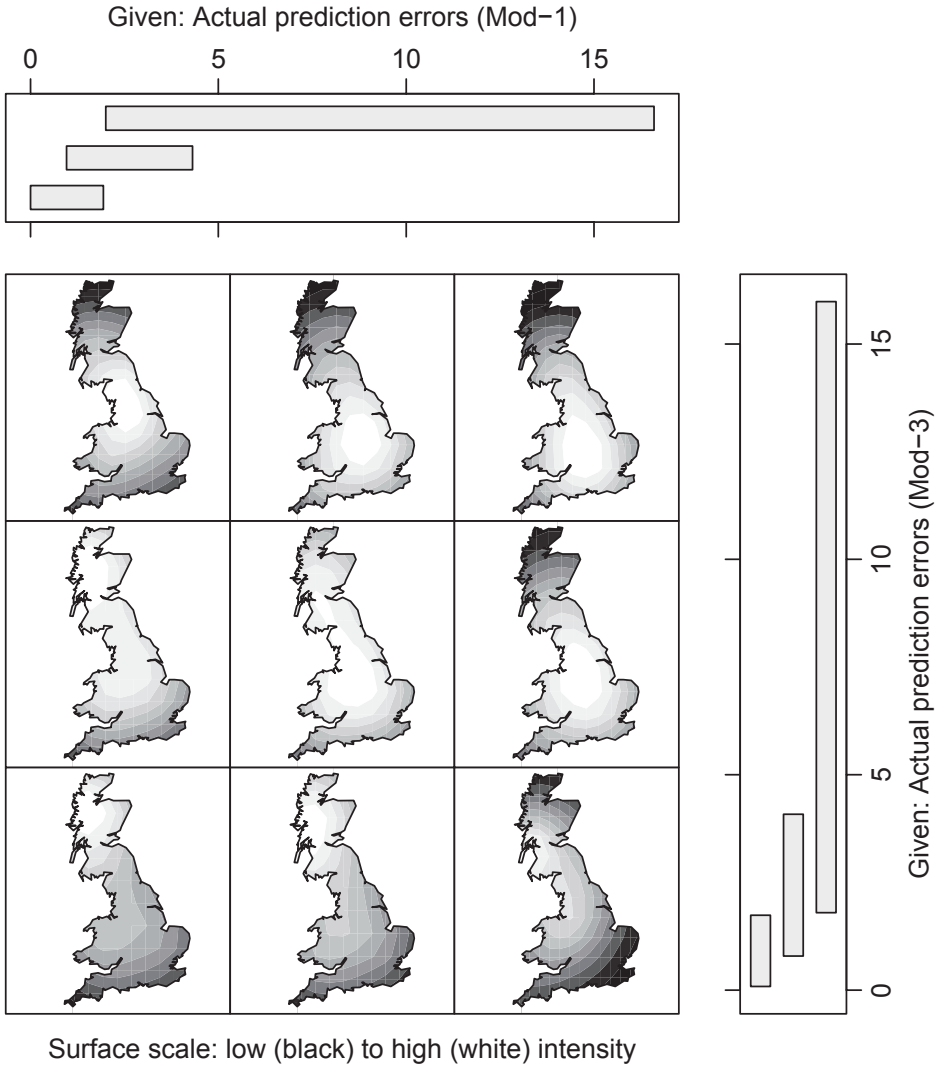

Figure 6. Critical load comap of the actual (absolute) prediction errors, $\left| z(\mathrm{x}_v) - \hat{z}(\mathrm{x}_v) \right|$ from Mod-1 compared with those from Mod-3; at the validation sites.

5.3. Bivariate visualisations of MWK outputs with comaps

We now provide examples of bivariate visualisations of MWK outputs using comaps for each of the three visualisation categories. First, we continue our model performance investigations with respect to the relationships between (i) the actual prediction errors for two different models and (ii) the actual and estimated prediction errors for a single model. As examples, a comap (using a 3×3 panel of map windows) is given for (i) Mod-1 and Mod-3 in Figure 6 and (ii) Mod-3 in Figure 7.

From Figure 6, models predict equally as well in areas of clustering intensity along the three leading diagonal map windows. Thus, prediction accuracy is relatively good for both models in NW Scotland (the bottom-left map window), but relatively poor

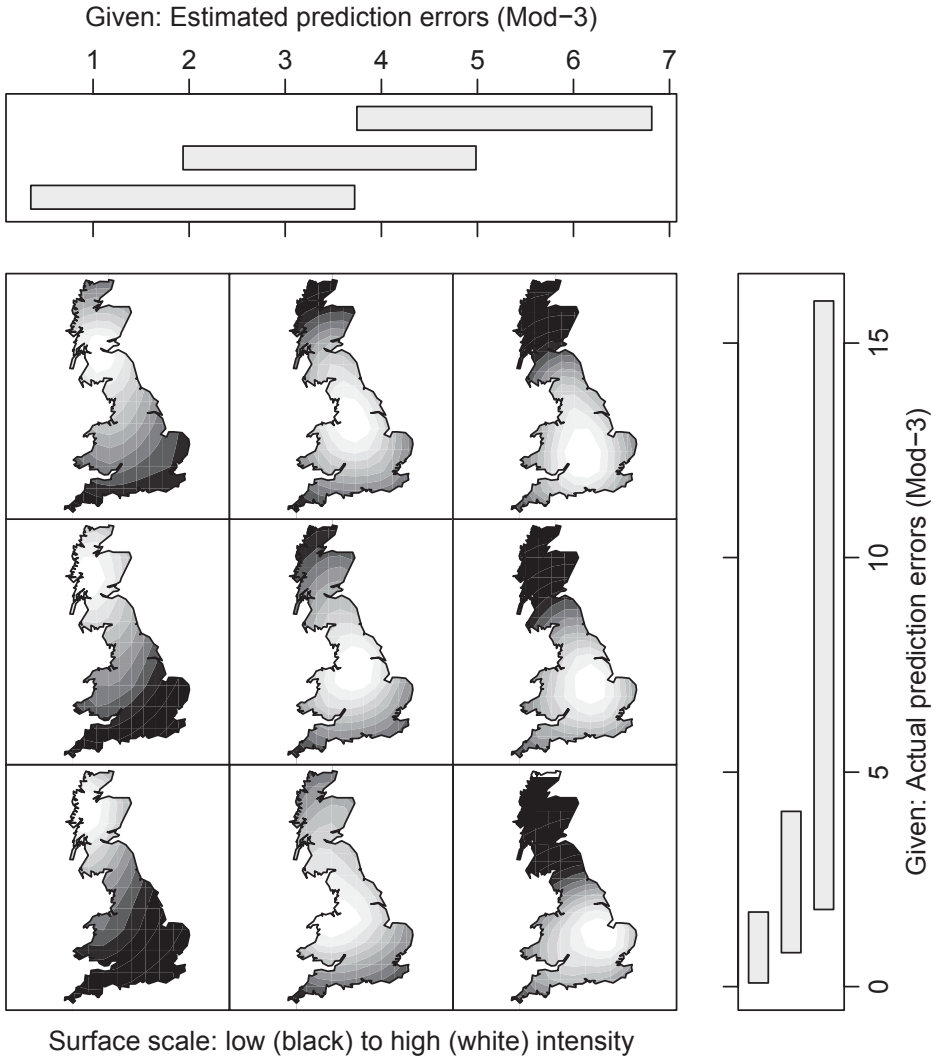

Figure 7. Critical load comap from Mod-3 output data: the estimated prediction standard errors, $\sigma(\mathrm{x}_v)$ and the actual (absolute) prediction errors, $\left| z(\mathrm{x}_v) - \hat{z}(\mathrm{x}_v) \right|$.

for both models in Central England (the top-right map window). The robust model (Mod-3) predicts much better than the basic model (Mod-1) in areas of clustering intensity found in the bottom-right map window. As expected from the univariate visualisations, this improved performance occurs on the coast of NW England, but now we also see an improved performance on the coast of south-west (SW) Scotland. Conversely, Mod-3 predicts less well than Mod-1 in other areas of Northern England (see top-left map window).

From Figure 7, we can identify where the estimated prediction errors have a good local correspondence with the actual prediction errors for Mod-3. These regions of clustering intensity are depicted along the leading diagonal map windows. For example, small estimated prediction errors tend to coincide with small actual predictions errors in NW

Scotland, as shown in the bottom-left map window. Alternatively, estimated prediction errors tend to over-estimate the actual prediction errors in an area centred near the coast of Eastern England (bottom-right map window). Clearly, the comaps in Figures 6 and 7 provide a distinct improvement over the univariate visualisations in Figure 4, as we are most interested in joint investigations of model performance data, rather than two separate ones. Comaps reduce subjectivity in identifying regions of data similarity and dissimilarity, enabling a more informed model choice.

For model specification, we use a comap (with a 3×4 panel of map windows) to investigate the pairwise relationship between the localised number of winsorised data edits (for Mod-3) and the localised number of adjusted boxplot outliers (for Mod-4). Here the comap in Figure 8 is constructed such that the fourth column reflects locations where the number of data edits always exceeds the number of boxplot outliers. From Figure 8, it is difficult to find a coherent pattern in the relationship between these data. Intuitively, relationships are expected and a re-specification of our robust models may provide them. That is, we could experiment with different values of the parameter b, which controls the number of data edits in Mod-3 and/or re-specify the outer fences of the boxplot used in Mod-4 (see Section 3.3). Alternatively, we could re-specify Mod-4, such that a robust prediction is only applied if the number of data edits found at each target location in Mod-3 exceeds some threshold (i.e. abandon the use of the boxplot diagnostic for this purpose altogether). In experimentation, Mod-4 is re-specified in this manner where a robust prediction is only applied at locations where the number of data edits exceeded eight (reflecting the fourth column of the comap in Figure 8). This re-specified model results in a slight improvement in performance (MPE = –0.03, RMPSE = 4.40, MAPE = 3.07, MSDR = 1.479, ERR-CORR = 0.43, G-STAT = 0.969, M-PCI-W = 6.17). Evidently, this particular model specification requires further work, but we have demonstrated that a comap can be useful in flagging a specification issue that requires additional scrutiny.

For model parameterisation, we can use comaps to investigate the local variographic relationship between the correlation range $a(\mathbf{x})$ and the relative structural variability RSV($\mathbf{x}$) for a given model. Such comaps can identify regions most suited to kriging, as kriging performs well in a relative sense when RSV($\mathbf{x}$) values tend to 100% together with a long correlation range. In essence, we can use comaps to identify regions where spatial dependence is relatively strong or weak. As an example, we provide a comap of this local variogram parameter data for Mod-1 in Figure 9, where it appears that spatial dependence is strong in an area that traverses the border of Wales and England (top-right map window), but weak over much of Scotland (bottom-left map window). Observe that this comap taken together with the surfaces of $c_0(\mathbf{x})$ (Figure 5b and d) provides clear evidence of variogram non-stationarity.

Finally, it is interesting to identify regions where local variographic structure changes when a robust MWK model is used. Intuitively, it is expected that spatial structure would tend to be strengthened with a robust model, as the influence of outliers on the local variography has been minimised. As an example, we present a comap of RSV($\mathbf{x}$) data from Mod-1 and from Mod-3 in Figure 10 (i.e. we now use RSV($\mathbf{x}$) on its own as a guide to the strength of local spatial structure). First, if we look at the axes of the comap, the RSV($\mathbf{x}$) data for the robust model (Mod-3) are indeed generally higher than those found for the basic model (Mod-1); however, there are regional differences. For example, RSV($\mathbf{x}$) data for Mod-3 are lower than those found for Mod-1 in a small region of NW Scotland (bottom-right map window).

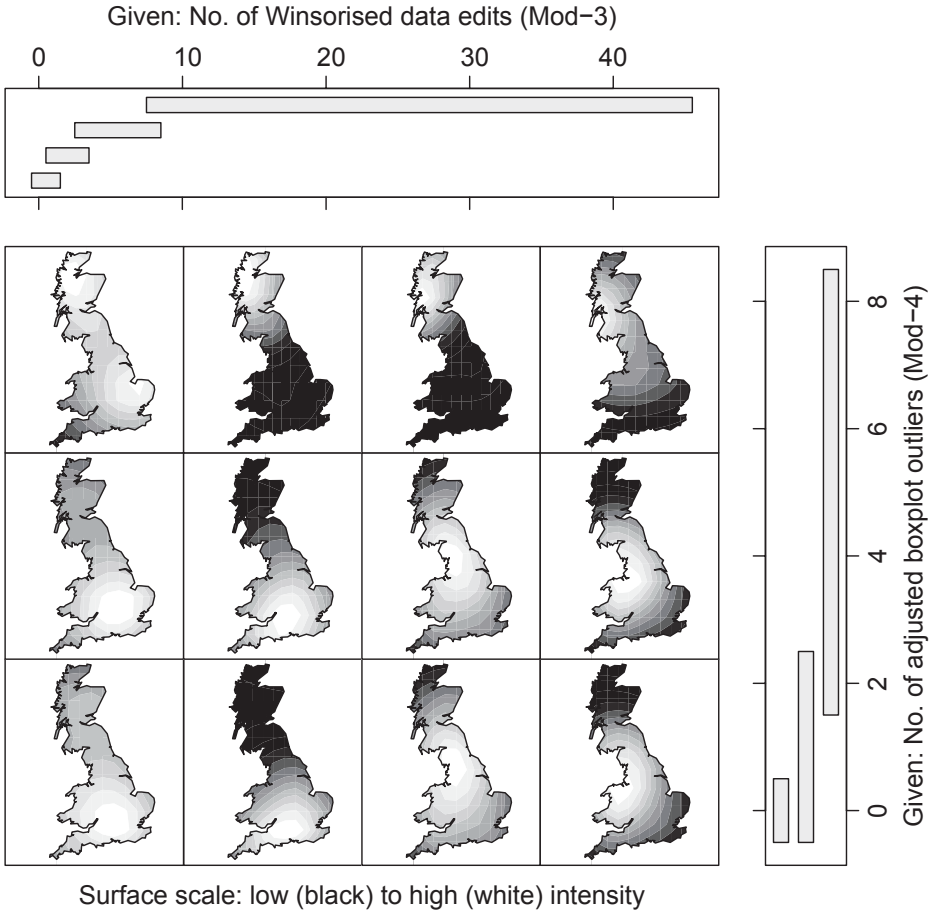

Figure 8. Critical load comap for the number of winsorised data edits found in each local calibration data set (for Mod-3) and the number of adjusted boxplot outliers found in each local calibration data set (for Mod-4) with respect to validation sites.

6. Case study B: house prices

Our second case study data is a set of house prices in London. Data were obtained for 11,285 geo-coded properties in London during 1998 from the Nationwide Building Society. These data were used to generate a sample of 2809 non-overlapping observations, which in turn was split into model calibration and validation data sets of 1405 and 1404 observations, respectively (Figure 2b). Again this data set suits a non-stationary model, where preliminary investigations provided evidence of (i) positively skewed data, (ii) local trends (Figure 11a and b), (iii) different levels of local variation (Figures 2b, 11c and d) and local skewness, (iv) spatial outliers, (v) a nested global variogram whose structure at lower lags is strongly affected by outliers (Figure 2d) and (vi) structure in the local variography not only varying across space but also varying according to the influence of outliers. GW summary statistic surfaces for these data are given in Figures 11a, b, c and d comparing non-robust with robust versions. Local differences in these paired surfaces are taken to confirm the presence of a skewed distribution with outliers. This behaviour in house price is not unexpected, as house prices can change markedly from one street to

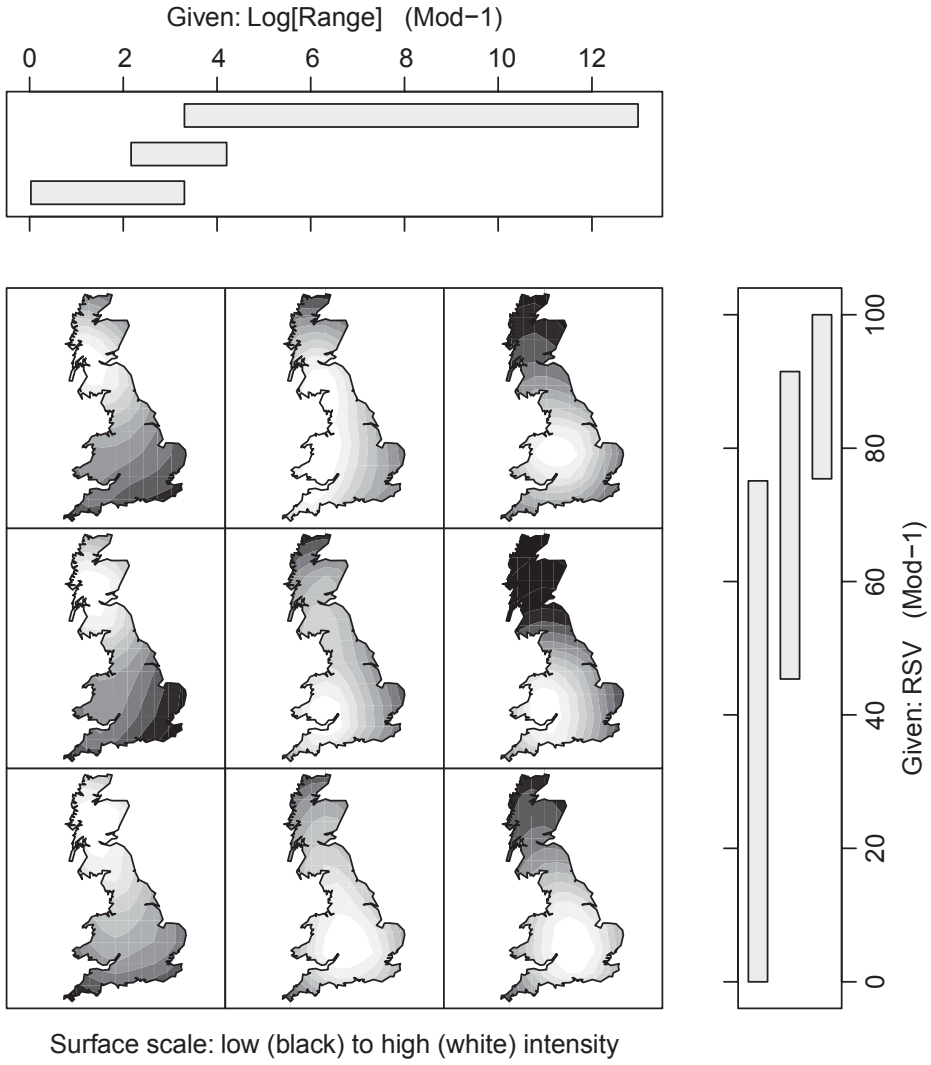

Figure 9. Critical load comap from Mod-1 output data: the local correlation ranges (in logs) and the local RSV values at the validation sites.

the next, which also gives rise to a high nugget effect in the variogram (see the classic estimator in Figure 2d).

6.1. Model performance at the validation sites: standard diagnostics

Model performance diagnostics using the house price data are presented in Table 2. Here we do not present the results for Mod-3 and Mod-4 as (i) results for Mod-3 are virtually identical to Mod-2, indicating no worth in a local Box–Cox transform, and (ii) results for Mod-4 are little different to Mod-2, indicating a universal spread of outliers across the study area (since Mod-3 and Mod-4 outputs are similar, but Mod-1 and Mod-4 outputs are strongly dissimilar). Clearly and unlike the first case study, the application of a robust model provides a distinct improvement over its non-robust counterpart in most aspects of

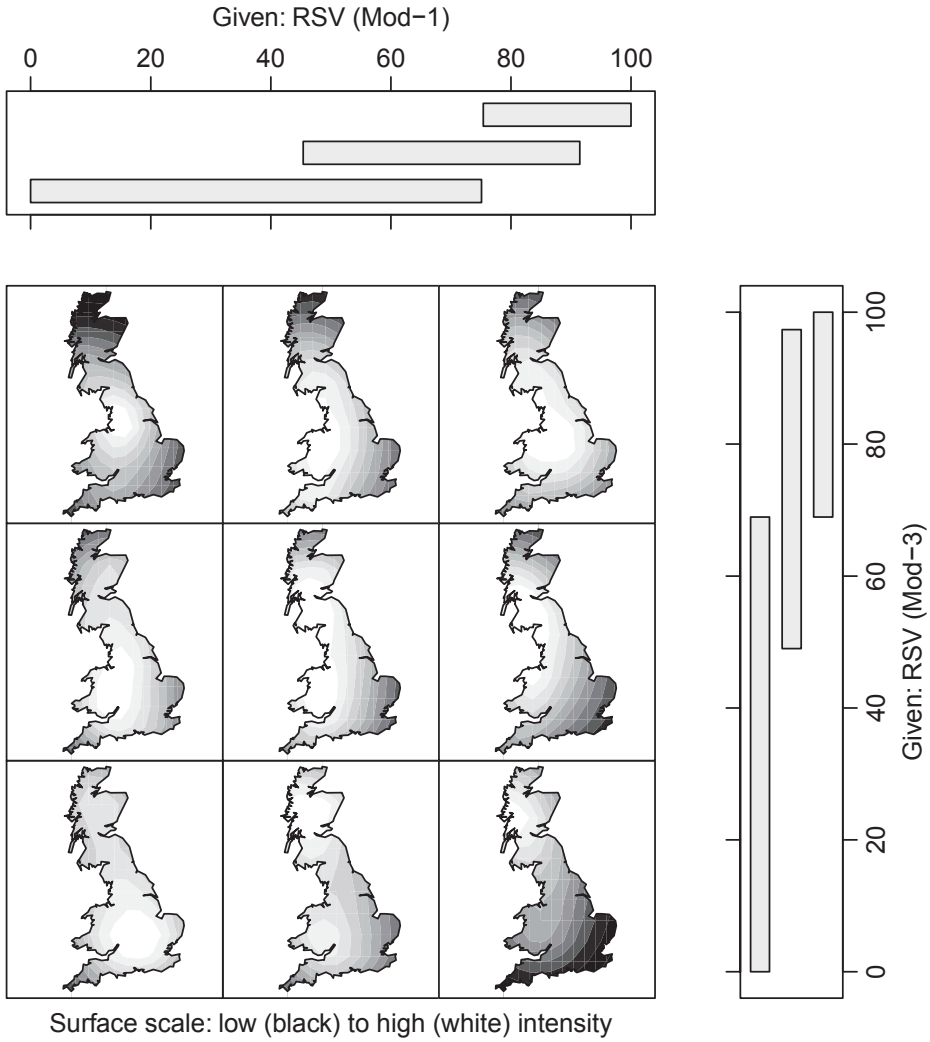

Figure 10. Critical load comap of the local RSV values from Mod-1 compared with those from Mod-3 at the validation sites.

model performance. However, for this case study, there appears little benefit in a local variogram kriging model in the first place, as Mod-2 specified with a global variogram (100% window size) performs in a similar manner to Mod-2 with a local variogram (where a 10% window size is considered optimal). This suggests a preponderance of proportional local variograms for these data (see Section 4.2). Partial evidence of this proportionality is provided in Figure 11e and f, where a clear linear relationship between local averages and local measures of variability exists. Complementary evidence is also provided by the fairly stationary CoV surface (Figure 2b). Thus, for this case study, we choose to focus our visualisations on the outputs of these best performing models (Mod-2, 10% window size and Mod-2, 100% window size for variogram and prediction), together with the best performing non-robust model for context (Mod-1, 4% window size).

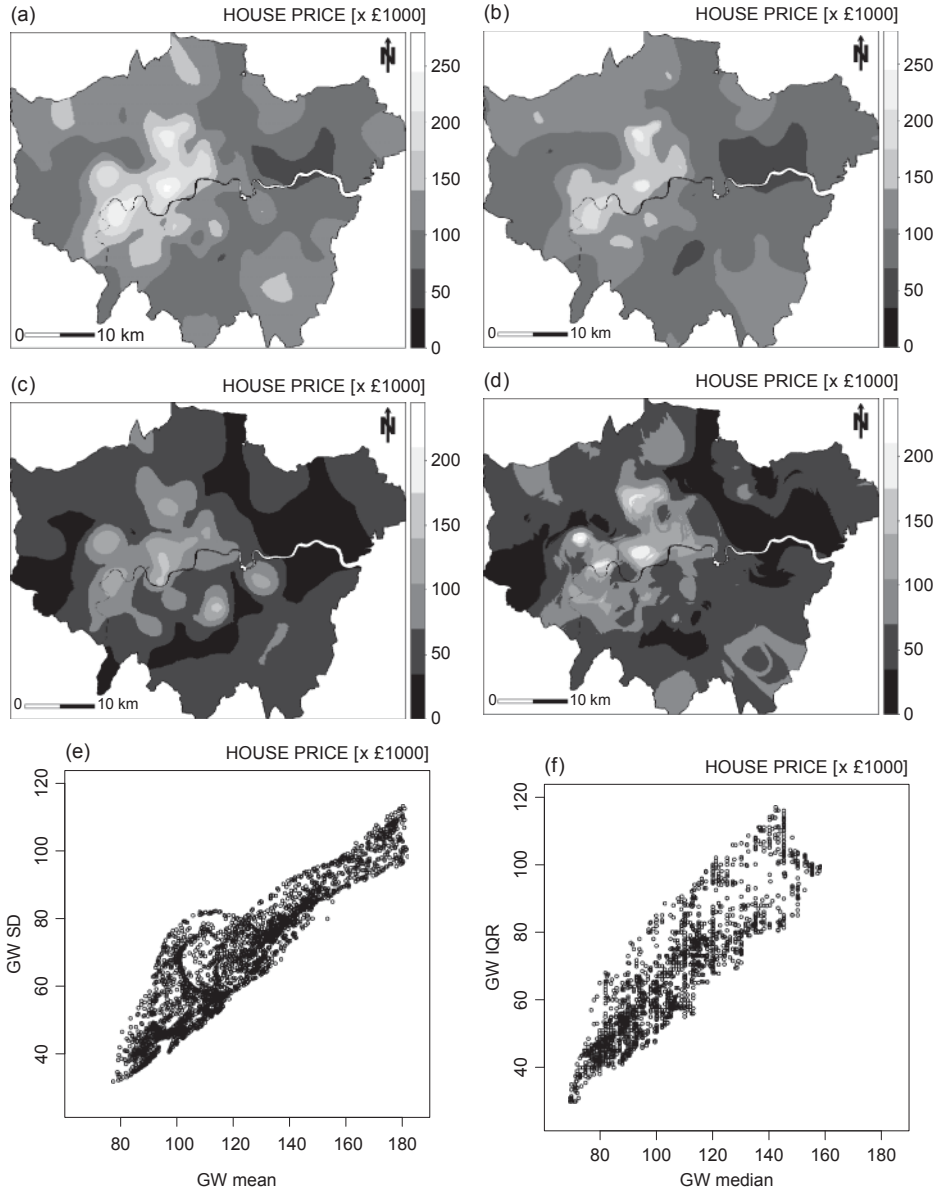

Figure 11. GW summary statistic surfaces for house price data: (a) mean; (b) median; (c) standard deviation (SD) and (d) interquartile range (IQR). Scatterplot relationships for the house price data between (e) GW means and SDs and (f) GW medians and IQRs. All GW statistics found using the full data set with bandwidths set at 2.5% and 15% for the surfaces and scatterplots, respectively.

6.2. Bivariate visualisations of MWK outputs with comaps

We limit our visualisations to an investigation of model performance, using only comaps. Furthermore, we demonstrate the value of using GW versions of the G-STAT and M-PCI-W diagnostics for this purpose. These diagnostics are found using a box-car kernel and provide an interesting and novel way of investigating model performance locally.

Table 2. Prediction and prediction uncertainty accuracy at the validation sites (house price data).

Model	Window size: variogram (%)	Window size: prediction (%)	MPE $\times 10^{-3}$	RMSPE $\times 10^{-3}$	MAPE $\times 10^{-3}$	MSDR	ERR-CORR	G-STAT	M-PCI-W
Ideal	–	–	0	$\rightarrow 0$	$\rightarrow 0$	1	$\rightarrow 1$	1	$\rightarrow 0$
NN10	–	–	−0.68	96.29	61.21	1.618	−0.04	0.929	120.69
Mod-1	100[a]	100	−3.11	61.11	42.21	0.664	−0.01	0.865	119.81
	100[a]	5	−1.10	60.51	41.11	0.632	−0.02	0.850	121.53
	30	30	−4.25	64.90	45.36	0.616	0.23	0.874	125.63
	20	20	−1.37	63.49	43.32	0.658	0.24	0.875	119.18
	10	10	−2.74	63.37	44.01	0.819	0.24	0.903	113.25
	5	5	−2.05	62.07	42.64	0.868	0.28	0.909	109.18
	4	**4**	**−1.71**	**61.89**	**42.10**	**0.868**	**0.28**	**0.907**	**109.42**
Mod-2	**100[a]**	**100**	**4.50**	**61.37**	**40.09**	**1.685**	**0.32**	**0.988**	**68.25**
	100[a]	5	3.23	60.89	40.04	1.671	0.32	0.989	71.22
	30	30	2.69	60.68	40.24	1.339	0.34	0.970	76.75
	20	20	3.06	60.86	40.29	1.465	0.33	0.976	75.47
	10	**10**	**3.12**	**60.82**	**40.21**	**1.556**	**0.33**	**0.983**	**74.09**
	5	5	2.62	61.04	40.53	1.673	0.32	0.985	74.74
	4	4	2.50	60.93	40.44	1.612	0.32	0.985	74.96

Notes: Models highlighted in bold are chosen for further scrutiny. Window sizes smaller than 4% suffered from calibration difficulties. [a]These naive models are the equivalent stationary (global) variogram kriging models.

As these diagnostics are intrinsically linked to each other, it is ideal to visualise them using comaps. The specification of a box-car kernel entails that the local G-STAT and M-PCI-W values are calculated simply using only nearby data from some target location. These local data subsets consist of the actual data $z(\mathbf{x}_v)$, the MWK predictions $\hat{z}(\mathbf{x}_v)$ and their standard errors $\sigma(\mathbf{x}_v)$.

Comaps of local G-STAT and M-PCI-W data, for each of our three retained models, are given in Figures 12–14 (specified with a 4×4 panel of map windows). Here we need to focus our attention on the off-diagonal map windows, where areas of strong model performance relate to clustering intensity in the bottom-right map window, whereas areas of weak model performance relate to clustering intensity in the top-left map window. Although our most accurate models (Mod-2 with a global variogram and Mod-2 with a local variogram) perform similarly in a global sense (from Table 2), this similarity does not always extend locally. For example, Mod-2 with a global variogram performs strongly in SE London, whereas Mod-2 with a local variogram not only performs well in the same area but also in an area of NW London. Many local differences are also observed between the robust models and the non-robust model (Mod-1 with a local variogram), and in this respect, one model could be preferred according to its superior performance in a key real estate area of interest. From the comap axes (for all three models), Mod-2 with a local variogram

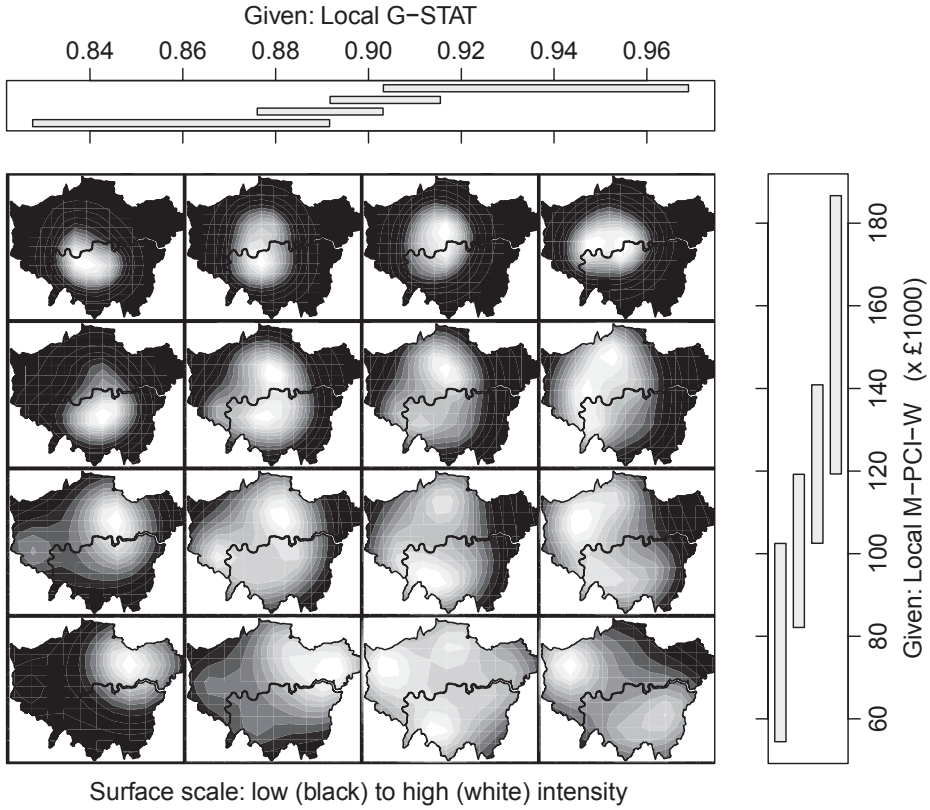

Figure 12. House price comap of local G-STAT versus local M-PCI-W ($\times$ £1000) values for Mod-1 (4% window size). G-STAT and M-PCI-W data are centred at the validation sites and were found by applying a box-car kernel with an adaptive bandwidth of 10% to the actual data, the predictions and their standard errors.

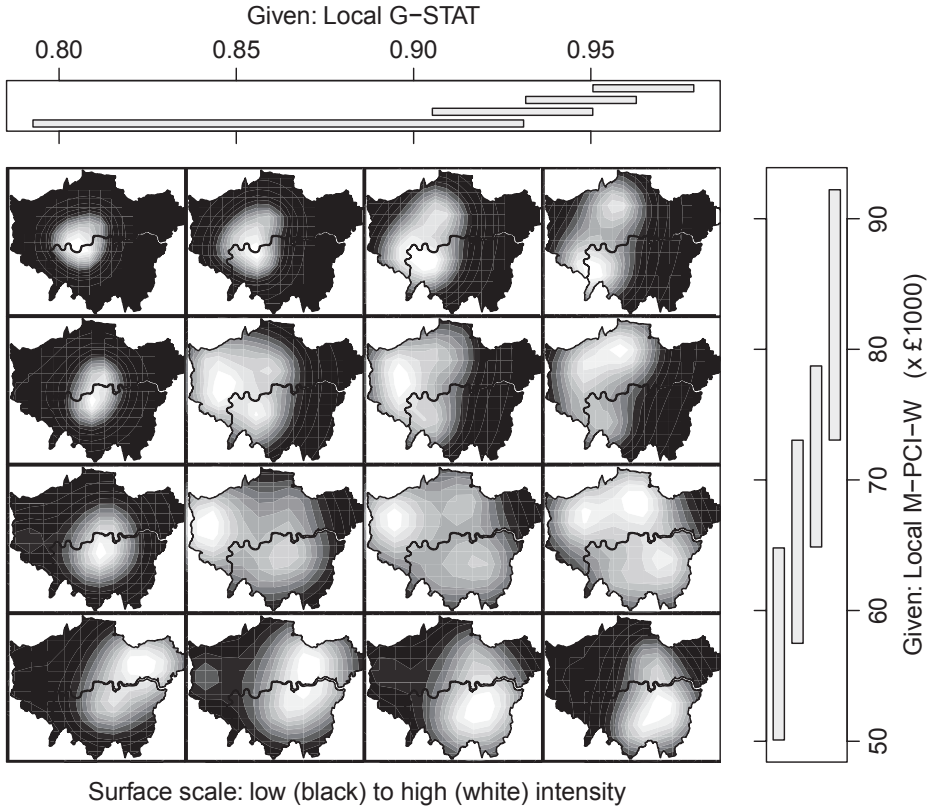

Figure 13. House price comap of local G-STAT versus local M-PCI-W (× £1000) values for Mod-2 (100% window size for variogram and prediction). G-STAT and M-PCI-W data were centred at the validation sites and were found by applying a box-car kernel with an adaptive bandwidth of 10% to the actual data, the predictions and their standard errors.

tends to have the highest G-STAT values and the lowest variability in this data, whereas Mod-2 with a global variogram tends to have the lowest M-PCI-W values and the lowest variability in these data. However, low M-PCI-W values are of little worth if coupled with low G-STAT values, and for this reason, Mod-2 with a local variogram is marginally preferred to Mod-2 with a global variogram. Clearly, our use of comaps with local G-STAT and M-PCI-W data has provided us with useful information that is hidden in the standard (more global) diagnostics.

7. Discussion and conclusions

In this study, we have shown how comaps have value in visualising paired outputs from a non-stationary spatial prediction model, both to help in its calibration and to help in its interpretation. In particular, we used comaps to investigate non-stationary variogram kriging models, where novel robust adaptations performed with promise. Bivariate visualisations via comaps, together with simple univariate visualisations and standard model diagnostics, each provided complementary insights into the performance, specification and parameterisation of our study models. Our use of comaps was particularly useful in

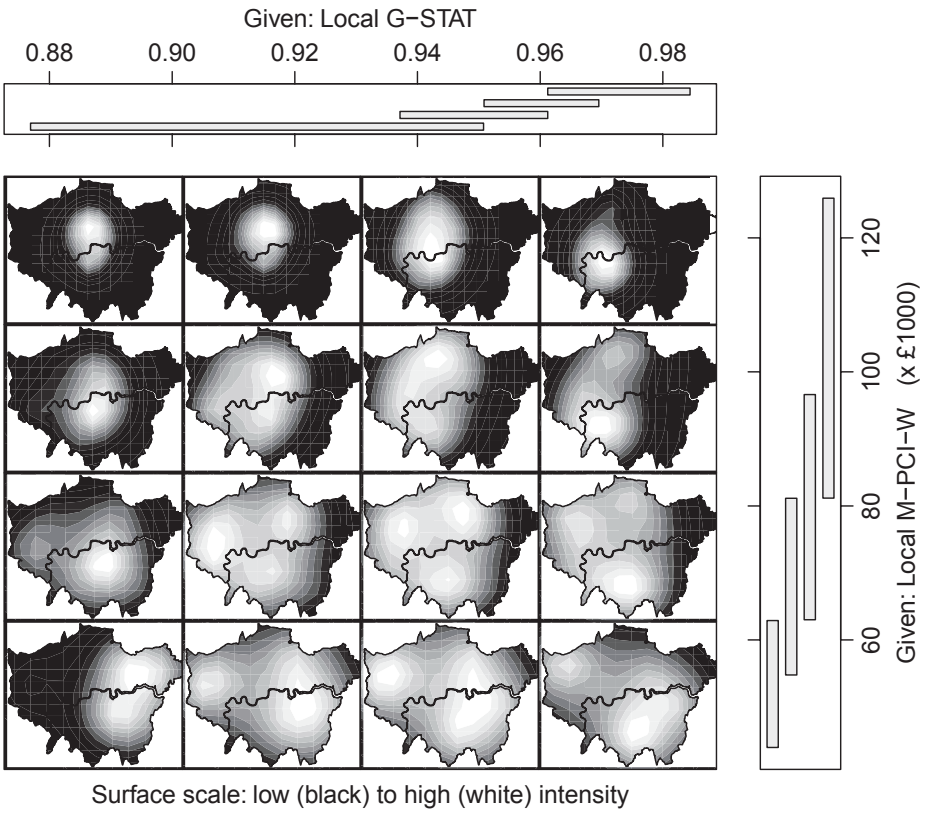

Figure 14. House price comap of local G-STAT versus local M-PCI-W (× £1000) values for Mod-2 (10% window size). G-STAT and M-PCI-W data were centred at the validation sites and were found by applying a box-car kernel with an adaptive bandwidth of 10% to the actual data, the predictions and their standard errors.

the development of the robust models whose benefits required much scrutiny. We also introduced a novel local assessment of PCI accuracy in conjunction with a comap, which should be of value to any spatial predictor and not only the non-stationary ones investigated here.

Further work could investigate our models within an interactive visualisation environment. For example, the variographic methods of Glatzer and Müller (2004) are useful, or more generally, the methods demonstrated in Demšar *et al*. (2008) in a non-stationary regression context would be transferable. As all our predictors are univariate in construction, we did not make use of available explanatory data. However, the extension to the multivariate case is straightforward, which for a robust form would represent a novel adaptation. For the critical load data, such predictors would complement those found in Harris and Juggins (2011). For the house price data, such predictors are commonly informed with *hedonic* variables (Páez 2009).

Acknowledgements

This article was funded by a Strategic Research Cluster grant (07/SRC/I1168) of the Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

Notes

1. All study algorithms were implemented in R (Ihaka and Gentleman 1996) using adaptations of existing functions. Plots similar to comaps can be found in the GeoDa program (Anselin *et al*. 2006). Software to implement basic MWK models can be found in the VESPER program (Whelan *et al*. 2002) and in the 'Geostatistical Analyst' extension for ArcGIS 10 (ESRI Corp., Redlands, CA, USA).
2. As an alternative to comaps, GW correlation coefficient surfaces could be found to investigate paired variables. However, GW correlations cannot distinguish between relationships that vary linearly across space, even if values are not close to one another – and as such, important information can be concealed. For this reason, we do not promote the use of GW correlations in this context.
3. As the uncertainty of the estimated SK mean will be larger for small window size predictions than for large window size predictions, it is important that this uncertainty is accounted for in our MWK models. We do this by calculating the variance of the estimation error $\mathrm{Var}\{\hat{m}(\mathbf{x}_0) - m(\mathbf{x}_0)\}$ and then add this variance to the SK variance at location $\mathbf{x}_0$. This adjustment provides the correct evaluation of uncertainty of the SK prediction. The variance is found using $\mathrm{Var}\{\hat{m}(\mathbf{x}_0) - m(\mathbf{x}_0)\} = \bar{C}(\mathbf{x}_0, \mathbf{x}_0) + \sum_{i=1}^{N}\sum_{j=1}^{N} \psi_i \psi_j C(\mathbf{x}_i, \mathbf{x}_j) - 2\sum_{i=1}^{N} \psi_i \bar{C}(\mathbf{x}_i, \mathbf{x}_0)$, where N is the prediction window size (i.e. the number of data used to estimate the SK mean), ψ_i is the weight assigned to the ith datum (i.e. $\psi_i = 1/N$ as the arithmetic mean is used), $\bar{C}(\mathbf{x}_0, \mathbf{x}_0) = \int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S} C(\mathbf{u}, \mathbf{v})\, d\mathbf{u}\, d\mathbf{v}$ and $\bar{C}(\mathbf{x}_i, \mathbf{x}_0) = \int_{\mathbf{v}\in S} C(\mathbf{x}_i, \mathbf{v})\, d\mathbf{v}$, where S is the prediction window inside which the SK mean is estimated and $\mathbf{x}_0$ an arbitrary point for denoting the window S (i.e. its centroid). These integrals are approximated by summations in a similar manner as that done in block kriging.
4. Commonly, λ is estimated with an assumption of uncorrelated data. To overcome this simplicity, maximum likelihood techniques to estimate λ while simultaneously accounting for correlated data can be found in Kitanidis and Shen (1996) and Christensen *et al*. (2001), both of which can suffer from identification issues with respect to a coherent estimation of the transform, trend and variogram parameters. Mindful of these difficulties and mindful of increasing MWK model complexity, we estimate λ as follows: (i) find an initial estimate of λ using maximum likelihood with an assumption of uncorrelated data; (ii) find a revised estimate of λ using maximum likelihood, but hold fixed variogram parameters found from the transformed data of (i). Thus, the likelihood is still a function of λ only, but λ is estimated with some consideration that the data are correlated.
5. A robust version of the MSDR statistic is promoted in Lark (2000) and Marchant *et al*. (2010).
6. ERR-CORR is not expected to be very strong as $\sigma(\mathbf{x}_v)$ is a statistical measure that is not going to be coincidental with $|z(\mathbf{x}_v) - \hat{z}(\mathbf{x}_v)|$. However, for a non-stationary variogram predictor, this correlation is expected to be stronger than that found in the stationary variogram case. For the stationary case, and referring to the method of Verly in David (1988), on average and assuming a Gaussian distribution for the errors, 75% of the data will be under the 1:1 line of the plot of $|z(\mathbf{x}_v) - \hat{z}(\mathbf{x}_v)|$ versus $\sigma(\mathbf{x}_v)$.
7. Values of MSDR, ERR-CORR, G-STAT and M-PCI-W are reported for the NN10 model. Here we find the NN10 prediction variance at any location $\mathbf{x}$ using the same expression as that used for the SK variance (Schabenberger and Gotway 2005, p. 224). This prediction variance is only a function of the variogram and the weights used in the linear predictor. Thus, to evaluate the uncertainty of the NN10 prediction, the parameters from a raw data (global) variogram are required and the SK weights are replaced with the single weight of 1 used in the NN10 prediction.

References

Anselin, L., Ibnu, S., and Youngihn, K., 2006. GeoDa: an introduction to spatial data analysis. *Geographical Analysis*, 38, 5–22.

Box, G.E.P. and Cox, D.R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society B*, 26, 211–252.

Brunsdon, C., 2001. The Comap: exploring spatial pattern via conditional distributions. *Computers, Environment and Urban Systems*, 25, 53–68.
Brunsdon, C., Corcoran, J., and Higgs, G., 2007. Visualising space and time in crime patterns: a comparison of methods. *Computers, Environment and Urban Systems*, 31, 52–75.
Brunsdon, C., Fotheringham, A.S., and Charlton, M.E., 2002. Geographically weighted summary statistics: a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems*, 26, 501–524.
Case, B., *et al*. 2004. Modeling spatial and temporal house price patterns: a comparison of four models. *Journal of Real Estate Finance Economics*, 29, 167–191.
Cattle, J.A., McBratney, A.B., and Minasny, B., 2002. Kriging methods evaluation for assessing the spatial distribution of urban soil lead contamination. *Journal of Environmental Quality*, 31, 1576–1588.
Chilès, J.P. and Delfiner, P., 1999. *Geostatistics: modelling spatial uncertainty*. New York: Wiley.
Christensen, O.F., Diggle, P.J., and Ribeiro, P.J., 2001. Analysing positive-valued spatial data: the transformed Gaussian model. *In*: P. Monestiez, D. Allard, R. Friodevaux, eds. *GeoENV III – geostatistics for environmental applications*. Dordrecht: Kluwer Academic Publishing.
CLAG-Freshwaters 1995. *Critical loads of acid deposition for United Kingdom freshwaters*, Sub-report on Freshwaters. ITE Penicuik: Critical Loads Advisory Group.
Cleveland, W.S., 1993. *Visualising data*. Summit, NJ: Hobart Press.
Costa, J.F., 2003. Reducing the impacts of outliers in ore reserves estimation. *Mathematical Geology*, 35, 323–343.
Cressie, N. and Hawkins, D.M., 1980. Robust estimation of the variogram. *Mathematical Geology*, 12, 115–125.
David, M., 1988. *Handbook of applied advanced geostatistical ore reserve estimation*. Amsterdam: Elsevier.
Demšar, U., Fotheringham, A.S., and Charlton, M., 2008. Combining geovisual analytics with spatial statistics: the example of geographically weighted regression. *The Cartography Journal*, 45, 182–192.
Diggle, P.J., 1990. A point process modelling approach to raised incidence of a rare phenomenom in the vicinity of a pre-specified point. *Journal of the Royal Statistical Society A*, 53, 349–362.
Fotheringham, A.S., Brunsdon, C., and Charlton, M.E., 2002. *Geographically weighted regression: the analysis of spatially varying relationships*. Chichester: Wiley.
Glatzer, E. and Müller, W.G., 2004. Residual diagnostics for variogram fitting. *Computers & Geosciences*, 30, 859–866.
Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. *Geoderma*, 103, 3–26.
Haas, T.C., 1990. Lognormal and moving window methods of estimating acid deposition. *Journal of the American Statistical Association*, 85, 950–963.
Haas, T.C., 1996. Multivariate spatial prediction in the presence of non-linear trend and covariance non-stationarity. *Environmetrics*, 7, 145–165.
Haas, T.C., 2002. New systems for modelling, estimating, and predicting a multivariate spatio-temporal process. *Environmetrics*, 13, 311–332.
Harris, P., Charlton, M.E., and Fotheringham, A.S., 2010. Moving window kriging with geographically weighted variograms. *Stochastic Environmental Research and Risk Assessment*, 24, 1193–1209.
Harris, P. and Juggins, S., 2011. Estimating freshwater acidification critical load exceedance data for Great Britain using space-varying relationship models. *Mathematical Geosciences*, 43, 265–292.
Hawkins, D.M. and Cressie, N., 1984. Robust kriging: a proposal. *Mathematical Geology*, 16, 3–18.
Hubert, M. and Vandervieren, E., 2008. An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52, 5186–5201.
Ihaka, R. and Gentleman, R., 1996. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299–314.
Journel, A.G. and Huijbregts, C.J., 1978. *Mining geostatistics*. London: Academic Press.
Kitanidis, P.K. and Shen, K.-F., 1996. Geostatistical interpolation of chemical concentration. *Advances in Water Resources*, 19, 369–378.
Lahiri, S.N., Lee, Y., and Cressie, N., 2002. On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *Journal of Statistical Planning and Inference*, 103, 65–85.

Lark, R.M., 2000. A comparison of some robust estimators of the variogram for use in soil survey. *European Journal of Soil Science*, 51, 137–157.

Lloyd, C.D., 2010. Nonstationary models for exploring and mapping monthly precipitation in the United Kingdom. *International Journal of Climatology*, 30, 390–405.

Lloyd, C.D. and Atkinson, P.M., 2002. Nonstationary approaches for mapping terrain and assessing prediction uncertainty. *Transactions in GIS*, 6, 17–30.

Marchant, B., *et al*., 2010. Robust analysis of soil properties at the national scale: cadmium contents of French soils. *European Journal of Soil Science*, 61, 144–152.

Matérn, B., 1960. Spatial variation: stochastic models and their applications to problems in forest surveys and other sampling investigations. *Meddelanden fran Statens Skogforskningsinstitut*, 49, 1–144.

Nilsson, J. and Grennfelt, P., eds., 1988. *Critical loads for sulphur and nitrogen*. Copenhagen: Nordic Council of Ministers.

Páez, A. 2009. Recent research in spatial real estate hedonic analysis. *Journal of Geographical Systems*, 11, 311–316.

Páez, A., Long, F., and Farber, S., 2008. Moving window approaches for hedonic price estimation: an empirical comparison of modelling techniques. *Urban Studies*, 45, 1565–1581.

Pardo-Igùzquiza, E., Dowd, P., and Grimes, D., 2005. An automatic moving window approach for mapping meteorological data. *International Journal of Climatology*, 26, 665–678.

Reimann, C., Filzmoser, P., and Garrett, R., 2005. Background and threshold: critical comparison of methods of determination. *Science of the Total Environment*, 346, 1–16.

Schabenberger, O. and Gotway, C., 2005. *Statistical methods for spatial data analysis*. London: Chapman & Hall.

Tufte, E.R., 1990. *Envisioning information*. Cheshire, CT: Graphics Press.

Tukey, J.W., 1962. The future of data analysis. *The Annals of Mathematical Statistics*, 33, 1–67.

Wand, M. and Jones, C., 1995. *Kernel smoothing*. London: Chapman and Hall.

Whelan, B.M., McBratney, A.B., and Minansy, B., 2002. VESPER 1.5 – spatial prediction software for precision agriculture. *In*: P.C. Robert, R.H. Rust, and W.E. Larsen, eds. *Precision agriculture. Proceedings of the 6th international conference on precision agriculture*. Madison, WI: ASA/CSSA/SSSA.

Zhang, X., Eijkeren, J.C., and Heemink, A.W., 1995. On the weighted least-squares method for fitting a semivariogram model. *Computers & Geosciences*, 21, 605–608.