Geographically weighted methods and their use in network re-designs for environmental monitoring

13/01/14

Paul Harris[1]

Annemarie Clarke[2]

Steve Juggins[3]

Chris Brunsdon[4]

Martin Charlton[1]

[1] National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth, Ireland

[2] APEM Ltd, Llantrisant, UK

[3] School of Geography, Politics and Sociology, University of Newcastle, Newcastle upon Tyne, UK

[4] School of Environmental Sciences, University of Liverpool, Liverpool, UK

Corresponding author:

Paul Harris, National Centre for Geocomputation, Iontas Building, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

Tel: 00353 1 708 6731; Fax: 00353 1 708 6456; e-mail: Paul.Harris@nuim.ie

Abstract

Given an initial spatial sampling campaign, it is often of importance to conduct a second, more targeted campaign based on the properties of the first. Here a network re-design modifies the first one by adding and/or removing sites so that maximum information is preserved. Commonly, this optimisation is constrained by limited sampling funds and a reduced sample network is sought. To this extent, we demonstrate the use of geographically weighted methods combined with a location-allocation algorithm, as a means to design a second-phase sampling campaign in univariate, bivariate and multivariate contexts. As a case study, we use a freshwater chemistry data set covering much of Great Britain. Applying the two-stage procedure enables the optimal identification of a pre-specified number of sites, providing maximum spatial and univariate/bivariate/multivariate water chemistry information for the second campaign. Network re-designs that account for the buffering capacity of a freshwater site to acidification are also conducted. To complement the use of basic methods, robust alternatives are used to reduce the effect of anomalous observations on the re-designs. Our non-stationary re-design framework is general and provides a relatively simple and a viable alternative to geostatistical re-design procedures that are commonly adopted. Particularly in the multivariate case, it represents an important methodological advance.

Keywords: non-stationarity; summary statistics; PCA; location-allocation; robust; acidification

1 Introduction

Given an initial spatial sampling campaign, it is often important to conduct a second sampling campaign based on the properties of the first. Here a network re-design modifies the first one by adding and/or removing sites so that maximum information is preserved. In this respect, we demonstrate the use of basic and robust geographically weighted (GW) methods combined with a location-allocation (L-A) algorithm (e.g. ReVelle and Eiselt 2005) as a means to design a second-phase sampling campaign in univariate, bivariate and multivariate contexts. We utilise outputs from GW summary statistics (GWSS) (Brunsdon et al. 2002; Fotheringham et al. 2002) and a GW principal components analysis (GWPCA) (Fotheringham et al. 2002; Harris et al. 2011), for input data into a L-A algorithm. Our re-design procedures are considered variations on that proposed by Kanaroglou et al. (2005), for the optimal location of urban air pollution monitors (but for a single variable only).

As a case study, networks of one, two and eight freshwater chemistry variables collected at 533 sites across Great Britain, are re-designed for a second, targeted sampling campaign of just 25 sites. This data was collected as part of a freshwater acidification study, for input into the calculation of associated critical load data (CLAG Freshwaters 1995). In this respect, weighted re-designs that account for the buffering capacity of a freshwater site to acidification (measured by its critical load) are also presented. Variability in the water chemistry data is mainly driven by the deposition of various (acidifying and non-acidifying) compounds, types of land use and geology; all of which vary both singly and in combination across Great Britain. Thus structures in the water chemistry data similarly vary across space; as shown in the companion study of Harris et al. (2014), also using GW methods.

Applying our re-design procedures enable optimal re-designs, with respect to the preservation of spatial and attribute information. The procedures are general and applicable to any environmental monitoring programme, where multiple variables are routinely measured. For freshwater environmental concerns, this not only includes re-designs for acidification (Hornung et al. 1995), but also for eutrophication (Pretty et al. 2003), both of which can benefit from targeted sampling where remedial actions are most urgent. As the collection of large environmental data sets continues (e.g. as required for the Water Framework Directive within the European Union; http://ec.europa.eu/environment/water/water-framework/) the availability of such re-design procedures will become increasingly valuable, because it would facilitate researchers and regulators to target areas where management actions may be required to prevent environmental degradation. Although, we present our network re-designs in univariate, bivariate and multivariate contexts, in practise a multivariate re-design is the most important, as once a sampling site is visited, it usually makes sense to measure multiple attributes. Our study is structured as follows: (i) literature review for spatial sampling; (ii) the network re-design methodology; (iii) a description of the study data; (iv) the re-design results; and (v) conclusions. All methods were implemented in R (http://www.r-project.org), many of which are available in the **GWmodel** R package (Lu et al. 2013). As a large number of abbreviations are used, the reader is helped with a list of abbreviations and their definitions (Table 1).

2 Background and context

Methods to conduct spatial sampling can be categorised into the following:

i. geometric-based (space-filling) (Royle and Nychka 1998);

ii. design- or probability-based with respect to random, systematic, stratified random, clustered and multi-stage sampling (e.g. de Gruijter et al. 2006);

iii. design-based accounting for key spatial effects such as autocorrelation (e.g. Griffith 2005; Webster et al. 2006) and heterogeneity (e.g. Wang et al. 2010);

iv. model-based via a (spatial) regression approach (Müller 2005; 2007);

v. model-based via a random fields approach - maximum entropy type (Caselton and Zidek 1984; Le and Zidek 1992; 2006);

vi. model-based via a random fields approach - geostatistical type (van Groenigen and Stein 1998; van Groenigen et al. 1999; Brus and Heuvelink 2007).

Reviews and linkages between the six categories can be found in Brus and de Gruijter (1997); Martin (2001); Xia et al. (2006); Wang et al. (2012). Observe that categories (iv-vi) relate to network re-designs, as methods require an initial data set to work with. Also a Gaussian process is commonly assumed for model-based methods, but other processes can be catered for. For categories (v-vi), the re-design can focus on: (a) prediction (e.g. McBratney et al. 1981; Ritter 1996), (b) (variogram) parameter estimation (e.g. Warrick and Myers 1987; Müller and Zimmerman 1999; Zhu and Stein 2005) or (c) a hybrid of both objectives (e.g. Zimmerman 2006).

For methods that stem from a geostatistical (kriging) perspective (category vi), the re-design is commonly optimised by the minimisation of the average kriging variance, where different optimisation algorithms are possible (Baume et al. 2011). Re-designs for one variable (say, using ordinary kriging) are common (e.g. van Groenigen et al. 1999), and include re-designs that are: (1) constrained by additional information via indicator kriging (e.g. van Groenigen et al. 2000) or (2) informed by covariates via regression kriging (Brus and Heuvelink 2007). Multivariate methods, where the re-design is for two or more variables simultaneously are rare; the cokriging approach of Vašát et al. (2010), aside. Important non-stationary extensions within the geostatistical framework address the severe short-fall of the kriging variance as a measure of local uncertainty (e.g. Journel 1986). For example, in a univariate re-design, Delmelle and Goovaerts (2009) weight the kriging variance with a local variance measure. Similarly, in a univariate re-design with covariates, Marchant et al. (2009) use regression kriging with a local (residual) variance term. Similarly, Haas (1992; 2002) specifies local variograms/cross-variograms for univariate and multivariate re-designs. For methods that stem from the maximum entropy framework (category v), univariate through to multivariate methods are possible, often incorporating important non-stationarities (Zidek et al. 2000). A drawback to these re-design methods is that they require considerable statistical expertise to fit and interpret their results. Furthermore, many methods are implicitly designed for space-time data, such as air pollution data (Zidek et al. 2000).

Our GW method-based re-designs are comparable to model-based methods in that our methods are for re-designing an existing network. Conceptually, they are closer to those from the geostatistical framework rather than the alternatives. In particular, they are comparable with non-stationary forms that account for local changes in variation (i.e. those described above, especially Haas (1992); Delmelle and Goovaerts (2009) in the univariate case; Haas (2002) in the bivariate/multivariate case). Our GW methods also link to design-based stratified sampling (category ii) with respect to an optimisation via prior information on within-strata variances; in that regions with high variability are sampled more intensively, than regions with low variability. A basic difference, however, between design-based methods and GW methods is that in the former, the selection probabilities are known. Furthermore, these known probabilities are used in making inferences.

Our methods directly build upon the univariate (and spatial-only) method of Kanaroglou et al. (2005), for locating urban air pollution monitors (for nitrogen dioxide, NO_2). Here a pragmatic and easily reproducible strategy was adopted, where a local variogram-type measure (weighted by population density) was used as *demand* data for input into an attendance maximising L-A algorithm (e.g. Holmes et al. 1972). The idea being that more monitors should be located in areas of high NO_2 variability and also where people actually live. We extend this method, in that our re-designs are possible for two or more variables, simultaneously. Essentially, we replace the local variogram-type measure with local variability outputs from univariate, bivariate and multivariate GW methods. The spatial scale of these local variability measures are, as far as possible, determined objectively, rather than subjectively. We also (non-geographically) weight these variability measures in the context of the study data (for weighted re-designs) and replace the attendance maximising L-A algorithm with a p-median L-A algorithm (e.g. Teitz and Bart 1968; Rosing et al. 1979). Our re-design framework is general and could optimally locate any (pre-specified) number of sample sites to any location - sampled or un-sampled in the first campaign (since the local variability measures can be found at any location). For this study however, we assume that only sites that were visited in the first campaign need to be chosen, and that a much reduced second campaign is required due to limited financial resources.

3 Methodology

In order to describe our re-design methodology, we proceed as follows: (i) a description of the p-median L-A algorithm; (ii) an overview of GW methods; (iii) the L-A demand data from GWSS and GWPCA methods; (iv) the rationale for choosing this demand data; (v) kernel bandwidth selection procedures for the GW methods; and (vi) a summary of the proposed methodology.

3.1 The p-median location-allocation (L-A) algorithm

L-A algorithms are routinely incorporated within a Geographical Information System where they are used to determine optimal locations of a service (say, hospitals, fire stations, schools, warehouses, etc). An L-A algorithm will match *supply* with *demand* via the use of objectives and constraints. The most common L-A algorithms are minimum distance (e.g. p-median) and maximum coverage (e.g. attendance maximising). The p-median algorithm will locate supply sites such that the sum of weighted distances for all demand sites from their nearest supply site is minimised. The distances are weighted by the demand data. More formally, the p-median algorithm minimises an objective function of the form

$$\boldsymbol{Z} = \sum_{\boldsymbol{i}=1}^{\boldsymbol{n_d}} \sum_{\boldsymbol{j}=1}^{\boldsymbol{n_s}} \boldsymbol{\lambda_i d_{ij} \delta_{ij}}\,, \tag{1}$$

where $\boldsymbol{n_d}$ is the number of demand sites; $\boldsymbol{n_s}$ is the number of potential supply sites; and $\sum \boldsymbol{\delta_{ij}} = \boldsymbol{n_p}$, where $\boldsymbol{n_p}$ is the number of supply sites to be located. The weight $\boldsymbol{\lambda_i}$ at site $\boldsymbol{i}$ represents demand; $\boldsymbol{d_{ij}}$ is the distance between sites $\boldsymbol{i}$ and $\boldsymbol{j}$; $\boldsymbol{\delta_{ij}}$ is the allocation decision variable (given a value of 1 if demand site $\boldsymbol{i}$ is served by a supply site in $\boldsymbol{j}$ and a value of 0 otherwise).

In context of our network re-designs, the supply sites will be those identified as potential sites for a second sampling campaign. In this case, we assume we have sufficient resources to re-sample at $\boldsymbol{n_p} = 25$ sites, selected from the $\boldsymbol{n_s} = 533$ sites of the initial water chemistry sampling campaign. The demand data will be local variability outputs from our GW methods, found at the same $\boldsymbol{n}_d = 533$ initial sites. The demand data can also be weighted for weighted re-designs (Sect. 5). Observe that we have chosen a p-median L-A algorithm. There are however, many L-A algorithms to choose from (ReVelle and Eiselt 2005), and one algorithm should be chosen that best suits the properties of the re-design process. Our reasons for choosing a p-median algorithm stem from its relative simplicity and reproducibility; it is possible that an alternative algorithm may be better suited to our case study data, but it is the over-arching concept of the re-design framework that we wish to focus on.

3.2 Geographically weighted methods

GW methods are primarily used to investigate spatial heterogeneity, where the form of the heterogeneity reflects the objective of the under-lying statistic or model (see Fotheringham et al. 2002; Harris et al. 2012). For example, a GW variance investigates spatial change in data variability (Brunsdon et al. 2002); a GW regression (GWR) investigates spatial change in response and predictor data relationships (see Brunsdon et al. 1998). A moving window weighting technique is used, where local models are calibrated at (sampled or un-sampled) locations (i.e. the window's centre). For an individual model at some calibration location, all neighbouring observations are weighted according to the properties of a distance-decay kernel function, and the model is locally-fitted to this weighted data. Thus the geographical weighting solely applies to the data in all GW methods, where each local model is fitted to its own GW data (sub)set. The size of the window over which this localised model might apply is controlled by the kernel's bandwidth. Small bandwidths lead to more rapid spatial variation in the results, while large bandwidths yield results increasingly close to the global model solution. Commonly, the exploration of spatial heterogeneity involves a simple non-stationarity test and a mapping of the outputs or parameters of the GW method. This may then direct a traditional (stationary) or sophisticated (non-stationary) analysis (e.g. when a rigorous inferential framework is required).

3.3 The L-A demand data from basic and robust GW methods

We utilise local variability outputs from basic and robust forms of GWSS and GWPCA, to use as demand data for input into a p-median L-A algorithm. Robust forms are specified, so as to reduce the effect of outliers on the network re-design (i.e. our definition of *robust* accords to the reviews of Rousseeuw et al. (2006) or Filzmoser and Todorov (2012)). This is important, as outliers can not only artificially increase local variability, but can also mask key features in local data structures. Robust procedures are possible for estimating location and scale in univariate, bivariate and multivariate contexts. Here, in the univariate case, we replace basic variance estimates with robust median absolute deviation (MAD) (Hampel 1974) estimates. In the bivariate and multivariate cases, robust estimates for the covariance matrix are needed and here, we specify the minimum covariance determinant (MCD) estimator (see Maronna et al. 2006). In particular, we find our demand data, via:

a. GW standard deviation (GWSD) values and GW median absolute deviation (GWMAD) values for basic and robust univariate re-designs, respectively;
b. Basic and robust GW correlation (GWCOR) values for bivariate re-designs;
c. Basic and robust GWPCA, where the demand data for the multivariate re-designs is the output that locally measures the percentage of the total variance in the original

data that is accounted for by a specified combination of components (i.e. via the component scores data). Call this demand data, PTV data.

GWSS

For attributes $\boldsymbol{x}$ and $\boldsymbol{y}$ at spatial location $\boldsymbol{i}$ where the geographical weights $\boldsymbol{w}_{ij}$ accord to a bi-square kernel weighting function (see section 3.5), definitions for a GWSD and a GWCOR, for a calibration sample size $\boldsymbol{n}$, are respectively $\boldsymbol{s}(\boldsymbol{x}_i) = \sqrt{\sum_{j=1}^{n} \boldsymbol{w}_{ij} \left(\boldsymbol{x}_j - \boldsymbol{m}(\boldsymbol{x}_i)\right)^2 \Big/ \sum_{j=1}^{n} \boldsymbol{w}_{ij}}$ and $\boldsymbol{\rho}(\boldsymbol{x}_i, \boldsymbol{y}_i) = \boldsymbol{c}(\boldsymbol{x}_i, \boldsymbol{y}_i) / \left(\boldsymbol{s}(\boldsymbol{x}_i)\boldsymbol{s}(\boldsymbol{y}_i)\right)$, where a local mean is $\boldsymbol{m}(\boldsymbol{x}_i) = \sum_{j=1}^{n} \boldsymbol{w}_{ij}\boldsymbol{x}_j \Big/ \sum_{j=1}^{n} \boldsymbol{w}_{ij}$, and where a local covariance is $\boldsymbol{c}(\boldsymbol{x}_i, \boldsymbol{y}_i) = \sum_{j=1}^{n} \boldsymbol{w}_{ij} \left\{\left(\boldsymbol{x}_j - \boldsymbol{m}(\boldsymbol{x}_i)\right)\left(\boldsymbol{y}_j - \boldsymbol{m}(\boldsymbol{y}_i)\right)\right\} \Big/ \sum_{j=1}^{n} \boldsymbol{w}_{ij}$. A definition for the MAD is $\boldsymbol{s}_{\mathrm{MAD}} = 1.4826 \cdot \mathrm{median}_i \cdot \left| \boldsymbol{x}_i - \mathrm{median}_k(\boldsymbol{x}_k) \right|$. A GWMAD involves finding local medians of both the sample data $\boldsymbol{x}_i$ and the expression $\left| \boldsymbol{x}_i - \mathrm{median}_k(\boldsymbol{x}_k) \right|$. Definitions for local medians are given in Brunsdon et al. (2002). For a robust GWCOR, we estimate the local covariance using the MCD estimator, whose objective is to find a subset of $\boldsymbol{h}$ observations whose basic sample covariance matrix has the lowest determinant. Crucial to the robustness and efficiency of this estimator is $\boldsymbol{h}$, and we specify a value of $\boldsymbol{h} = 0.75\boldsymbol{n}$, following Varmuza and Filzmoser (2009, p.43).

GWPCA

For GWPCA, a different localised PCA is computed at target locations**,** allowing a local identification of any change in the structure of a multivariate data set.

If spatial location $\boldsymbol{i}$ has coordinates $(\boldsymbol{u}_i,\boldsymbol{v}_i)$, then GWPCA involves regarding a vector of observed variables $\mathbf{x}_i$ as having a certain dependence on its location $\boldsymbol{i}$, where $\boldsymbol{\mu}(\boldsymbol{u}_i,\boldsymbol{v}_i)$ and $\boldsymbol{\Sigma}(\boldsymbol{u}_i,\boldsymbol{v}_i)$ are the local mean vector and the local covariance matrix, respectively. The local covariance matrix is $\boldsymbol{\Sigma}(\boldsymbol{u}_i,\boldsymbol{v}_i)=\mathbf{X}^{\mathrm{T}}\mathbf{W}(\boldsymbol{u}_i,\boldsymbol{v}_i)\mathbf{X}$, where $\mathbf{X}$ is the data matrix (with $\boldsymbol{n}$ observation rows and $\boldsymbol{m}$ variable columns); and $\mathbf{W}(\boldsymbol{u}_i,\boldsymbol{v}_i)$ is a diagonal matrix of geographic weights, generated using the bi-square kernel function. To find the local principal components at location $\boldsymbol{i}$, the decomposition of the local covariance matrix provides the local eigenvalues and local eigenvectors (or loading vectors) with $\mathbf{L}(\boldsymbol{u}_i,\boldsymbol{v}_i)\mathbf{V}(\boldsymbol{u}_i,\boldsymbol{v}_i)\mathbf{L}(\boldsymbol{u}_i,\boldsymbol{v}_i)^{\mathrm{T}}=\boldsymbol{\Sigma}(\boldsymbol{u}_i,\boldsymbol{v}_i)$**,** where $\mathbf{L}(\boldsymbol{u}_i,\boldsymbol{v}_i)$ is a matrix of local eigenvectors; $\mathbf{V}(\boldsymbol{u}_i,\boldsymbol{v}_i)$ is a diagonal matrix of local eigenvalues; and $\boldsymbol{\Sigma}(\boldsymbol{u}_i,\boldsymbol{v}_i)$ is the local covariance matrix. A matrix of local component scores $\mathbf{T}(\boldsymbol{u}_i,\boldsymbol{v}_i)$ can be found using $\mathbf{T}(\boldsymbol{u}_i,\boldsymbol{v}_i)=\mathbf{XL}(\boldsymbol{u}_i,\boldsymbol{v}_i)$**.** Thus at each data location for a GWPCA with $\boldsymbol{m}$ variables, there are $\boldsymbol{m}$ components, $\boldsymbol{m}$ eigenvalues, $\boldsymbol{m}$ sets of component loadings (with each set $\boldsymbol{m}\times\boldsymbol{m}$), and $\boldsymbol{m}$ sets of component scores (with each set $\boldsymbol{n}\times\boldsymbol{m}$). We can also obtain eigenvalues and their associated eigenvectors at un-sampled locations. For this study, we specifically require the local PTV data. For each component in turn, this data is found if we divide each local eigenvalue by $\mathrm{tr}(\mathbf{V}(\boldsymbol{u}_i,\boldsymbol{v}_i))$, and then multiply by 100. PTV data can be found at any location (sampled or not). For a robust GWPCA, we estimate the local covariance matrix using the MCD estimator, as specified with robust GWCOR.

3.4 Rationale for using GW methods for network re-design

1. For a univariate re-design, fewer sites should be selected in areas where data variation is low, reflected by low GWSD/GWMAD values. Conversely, more sites should be selected where data variation is high, reflected by high GWSD/GWMAD values.
2. For a bivariate re-design, fewer sites should be selected in areas where the correlation between the two variables is strong, reflected by a GWCOR tending to ± 1. Conversely, more sites should be selected where the correlation between the two variables is weak, reflected by a GWCOR tending to zero. Furthermore, for the sparse network of sites selected in areas with strong local correlations, one variable could be inferred from the other, entailing that only one variable needs sampling.
3. The rationale for the bivariate re-design directly extends to the multivariate re-design with GWPCA. Here we postulate that fewer sites should be selected in areas where the correlations (or collinearity) amongst the multivariate data are strong, reflected by high PTV outputs. In these areas, fewer sites are needed; and for sites that are selected, not all of the variables need to be sampled. Conversely, more sites should be selected in areas where the correlations amongst the data are weak, reflected by low PTV outputs. In these areas, more sites are needed and all variables carry important information (so all need to be sampled for).

3.5 Basic and robust bandwidth selection in the context of network re-design

For this study's GW methods, we generate the geographic weights $\boldsymbol{w}_{ij}$ using a bi-square kernel function, which can be defined as

$$\boldsymbol{w}_{ij} = \left(1 - \left(\boldsymbol{d}_{ij} / \boldsymbol{r}\right)^2\right)^2 \text{ if } \boldsymbol{d}_{ij} \leq \boldsymbol{r} \quad \boldsymbol{w}_{ij} = 0 \quad \text{otherwise,} \tag{2}$$

where the bandwidth is the geographic distance $\boldsymbol{r}$; and $\boldsymbol{d}_{ij}$ is the geographic distance between spatial locations of the $\boldsymbol{i}^{th}$ and $\boldsymbol{j}^{th}$ rows in the data matrix. The bandwidth can be specified as: (A) a *fixed* distance (where the number of local observations vary within the search window) or (B) an *adaptive* (varying) distance (where the number of local observations are fixed within the search window). For this study, we always specify the bandwidth as an adaptive distance, where the number of local observations is reported as a percentage of the full data set.

Crucial to any GW method is choosing the size of the bandwidth, so that the *true* scale of process heterogeneity is reflected in their outputs. In the context of network re-design, bandwidths that are too large will result in variability measures tending to their global form. Such demand data will have little influence in the L-A algorithm, resulting in a network re-design that only reflects the data's geometric properties (i.e. sites will be optimally and evenly dispersed across the initial design). Thus bandwidths need to chosen with care, as the re-design is strongly dependent on them. Bandwidths can be: (i) optimally-specified via a cross-validation procedure (provided an objective function exists); (ii) user-specified, but guided by the behaviour of some surrogate method that is expected to have similar properties; (iii) user-specified, guided by experience of the process under study; or (iv) user-specified, but guided by all three of the above methods. Of our study GW methods, only for GWPCA is cross-validation possible. For the rest, a surrogate procedure is used.

Bandwidths for GWSD and GWMAD data: univariate re-designs

For GWSD and GWMAD data, a surrogate procedure is used where we make use of the fact that local averages tend to scale with local variances in environmental data. This is known as the *proportional effect* in geostatistics (Chilès and Delfiner 1999). As such, it is reasonable to find bandwidths for GW means and GW medians, which can be found optimally, and then use the same bandwidths for calculating GWSD and GWMAD values, respectively. Basic and robust leave-one-out cross-validation procedures are followed, analogous to that used in GWR (see Brunsdon et al. 1998, and Farber and Páez 2007, respectively), but where local means/medians are used as predictors rather than local regressions. Robust procedures are specified as cross-validation procedures are themselves susceptible to outliers, even if the GW predictor is robust. A robust procedure down-weights the contributions of (residual) outliers when computing its goodness of fit (GOF) statistic. Thus for the basic procedure, the bandwidth $\boldsymbol{r}$ that minimises this GOF expression

$$\mathrm{RMSPE}(\boldsymbol{r}) = 1/\boldsymbol{n}\sqrt{\sum_{i=1}^{n}\{\boldsymbol{x}_i - \hat{\boldsymbol{x}}_{\neq i}(\boldsymbol{r})\}^2} \tag{3}$$

is taken as optimal, where RMSPE is the root mean squared prediction error and $\hat{\boldsymbol{x}}_{\neq i}(\boldsymbol{r})$ is the predicted value of $\boldsymbol{x}_i$ after the observation at $\boldsymbol{i}$ is removed. In the robust procedure, the bandwidth $\boldsymbol{r}$ that minimises this GOF expression

$$\mathrm{MAPE}(\boldsymbol{r}) = 1/\boldsymbol{n}\sum_{i=1}^{n}\left|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_{\neq i}(\boldsymbol{r})\right| \tag{4}$$

is taken as optimal, where MAPE is the mean absolute prediction error. Thus for univariate re-designs, we aim to find four surrogate bandwidths in total, two relate to GWSD data, whilst two relate to GWMAD data. The only re-design that is fully-robust is one that uses GWMAD demand data and where the GWMAD bandwidth relates to that found optimally for a GW median via the MAPE GOF data.

Bandwidths for basic and robust GWPCA: multivariate re-designs

We choose bandwidths for basic and robust GWPCA in accordance with an existing cross-validation procedure, where it is necessary to pre-specify (by experimentation) the number of retained components, $\boldsymbol{q}$ (where $\boldsymbol{q} < \boldsymbol{m}$, i.e. cannot specify all $\boldsymbol{m}$ components). A dual optimisation of both $\boldsymbol{r}$ and $\boldsymbol{q}$ is not currently viable, although work in this area is on-going. Again basic and robust procedures are possible, where GWPCA is viewed as a data model (Jolliffe 2002) and in doing so, enables leave-one-out GOF statistics to be found. The basic procedure uses the mean GOF, whilst the robust procedure uses the median GOF (i.e. analogous to those defined in expressions 3 and 4). For a given value of $\boldsymbol{q}$, the bandwidth $\boldsymbol{r}$ that minimises the specified GOF statistic is taken as optimal. Details are given in Harris et al. (2011). Thus we aim to find four optimal bandwidths in total, two relate to basic GWPCA, whilst two relate to robust GWPCA. The only re-design that is fully-robust is one that uses robust GWPCA PTV demand data and where the GWPCA bandwidth is found robustly. Observe that robust GWPCA is computationally intensive (due its use of the MCD estimator) and applying it within a leave-one-out cross-validation significantly increases this computational burden.

Bandwidths for basic and robust GWCOR data: bivariate re-designs

For GWCOR data, bandwidth selection is guided by a surrogate procedure. Here it is observed that an alternative bivariate re-design could be found by applying GWPCA with two variables ($\boldsymbol{m} = 2$), where an optimal bandwidth can only be found from $\boldsymbol{q} = 1$. This GWPCA re-design is expected to provide broadly similar results to that based on GWCOR, provided both methods are calibrated with the same bandwidth. Thus a surrogate bandwidth for GWCOR data is simply that found optimally for GWPCA, using the same two variables. This procedure is also useful in that a (simple) bivariate GWPCA re-design provides insight to the expected behaviour of the (more complex) multivariate GWPCA re-design (where $\boldsymbol{m} > 2$). That is, it can help to review our postulations of Sect. 3.4. Considering basic and robust forms are again needed, then for a bivariate re-design, we aim find four bandwidths in total. Two relate to basic GWCOR data, whilst two relate to robust GWCOR data.

3.6 Summary: key stages of the network re-design methodology

a. Decide on the potential locations for the second sampling campaign. In this study, we assume that only sites that were visited in the first campaign can be chosen. Thus all demand data sets are found at the $\boldsymbol{n} = 533$ study sites.

b. Determine the kernel bandwidths for GWSD data (basic), GWMAD data (robust) in the univariate re-design case; GWCOR data (basic and robust) in the bivariate case; and GWPCA (basic and robust) in the multivariate case.

c. From the specifications in (b), find all demand data sets for univariate, bivariate and multivariate re-designs. Map this data for context.

d. In the univariate case, the resultant GWSD and GWMAD data can be fed directly into the L-A algorithm. However, in the bivariate and multivariate cases, this demand data is needed: $1 - \left| \text{GWCOR} \right|$ and $100 - \text{PTV}$, respectively; each reflecting the nature of the postulations presented in Sect. 3.4.

e. Find the distance matrix for the study data, choose each demand data set in turn, and run the L-A algorithm with a pre-specified number of second campaign sites to optimally locate. In this case, 25 sites are taken to reflect the available resources.

f. Map the re-design results from each L-A run, indicating the level of demand at each of the chosen 25 sites. The L-A algorithm also *allocates* nearby sites to each chosen site (i.e. a spatial classification). This allocation could be mapped to provide insight into the results with respect to the initial sampling campaign, but is not strictly necessary in our network re-design context.

g. Include in the re-design results, 25 sites chosen when a demand data set consists of 533 equal values. This provides a benchmark re-design reflecting only the properties of the L-A algorithm, with respect to the sample configuration of the study data.

4 Case study: freshwater chemistry data for Great Britain

The study data is a subset of a water chemistry sampling programme for Great Britain; data that was used to calculate and map freshwater acidification critical loads (Kreiser et al. 1993). Sites were chosen to represent the most acid-sensitive water body within either a 10km (for sensitive areas) or 20km (for non-sensitive areas) grid square so that the minimum critical load is found. Data chosen for our study is composed of eight variables at 533 freshwater sites. These are: pH, alkalinity, conductivity, nitrate, sulphate, phosphate, total monomeric aluminium, and total organic carbon. All variables aside from pH were jointly transformed to approximate multivariate normality; and are thus named as: pH, Alk.T, Cond.T, NO3.T, SO4.T, PO4.T, AL.TM.T and TOC.T, respectively. For the GWPCA calibrations, this data is then standardised (thus covariance matrices are specified). Details on the selection and pre-processing of this data can be found in Harris et al. (2014).

Observe that data pre-processing decisions can strongly affect the local variability outputs of our GW methods, which can be further complicated in that the existence of outliers may be a key contributing factor in any differences observed. As a consequence, our network re-designs are dependent on the particular data pre-processing decisions we have taken. As our data are globally-transformed (and globally-standardised for GWPCA), there is no guarantee that the data will retain the associated properties at the scale of each local fit. The consequences of these decisions are most pertinent to GWPCA, as the standardisation should be conducted at the scale of each local PCA (local transforms are not so important). This local operation is not fully viable for GWPCA and here the use of globally-standardised data is considered a pragmatic alternative. A very *limited* GWPCA with locally-standardised data can be found however, and provides a check on this decision. Locally-standardised data (or at least an approximation to it) ensures that variables with the largest (local) variances do not dominate the local variability outputs and in turn, bias the network re-design. Observe that GWSD and GWCOR are already in a locally-standardised form, for our univariate and bivariate re-designs, respectively.

Variability measures for each variable are given in Table 2, where the largest SD and MAD values are for nitrate and phosphate. Basic and robust correlation matrices are given in Table 3, where the strongest correlations are between pH and alkalinity, which is to be expected. Weak correlations are also evident, while some variables, e.g. nitrate and aluminium are essentially uncorrelated. PCA eigenvalue and PTV data (Table 4) reveal that the first three components have eigenvalues greater than unity, where they collectively account for 79.9% or 81.6% of the variation in the data for basic and robust fits, respectively. These results reflect the relatively strong levels of collinearity amongst some of the variables. Basic and robust PCA scores data for the first component are mapped in Fig. 1. Spatial trends in this data tend to reflect where acidification is of most and of least concern; with high positive scores in areas of Scotland and Wales, and high negative scores in much of England, respectively. In general, and at this global scale, there is little difference between basic and robust forms. This does not imply however, that outliers will not have an influence, locally. Thus pursuing a robust GW methodology still has merit.

5 Results and discussion

In this section, we present a selection of network re-designs. In the univariate case, we choose pH as the variable for a targeted second sampling campaign. In the bivariate case, we choose aluminium and TOC. This particular variable pair is chosen, as they are likely to exhibit stronger local correlations in areas impacted by acidification, and so exhibit correlations that are not only different locally, but also different to that found globally. In the multivariate case, all eight variables are used.

5.1 Bandwidth selection

Bandwidths for the GW methods are given in Table 5, using the procedures described in Sect. 3.5. For GWSD and GWMAD data of the univariate re-designs, their surrogate bandwidths (via GW means and GW medians) are relatively small ranging from 3% to 7%, where one bandwidth function did not reach a minimum. Bandwidth functions for GW means and GW medians (for the basic selection procedure) are given in Figs. 2a-b, where in both cases, clear minimums are reached. For basic and robust GWCOR data of the bivariate re-designs, none of their surrogate bandwidth functions (via bivariate GWPCA) displayed a clear minimum and as such, bandwidths are chosen with judgement based on the behaviour of their function. Bandwidth functions for basic and robust bivariate GWPCA (for the basic selection procedure) are shown in Figs. 2c-d, where bandwidths are taken at local minimums found at 33% and 41%, respectively. Bandwidth functions using the robust selection procedure, behaved in a similar manner to that observed with the basic procedure.

For basic and robust GWPCA of the multivariate re-designs, GWPCA can be used directly to find its (optimal) bandwidth. However, we need to specify the number of $\boldsymbol{q}$ retained components. From the PCA results in Table 4, it is natural to proceed with $\boldsymbol{q} = 3$ or $\boldsymbol{q} = 4$ components as this provides cumulative PTV values of around 80% or around 90%, respectively. As such, we focus our GWPCA calibrations on bandwidth functions where $\boldsymbol{q} = 4$, as this corresponds to that of a reasonable PCA specification. Thus from Table 5, clear minimums were found at bandwidths of 48% and 56% for basic GWPCA using the basic and robust selection procedures, respectively. However, for both cases of robust GWPCA, the bandwidth function did not reach a minimum. The bandwidth functions for basic and robust GWPCA (for the basic selection procedure) are given in Figs. 2e-f, where the former reaches a minimum, whilst the latter does not.

As broadly similar bandwidths are found within each of the three groups of GW methods, we choose to use the same bandwidth for their respective re-designs. This is useful as it enables an objective comparison of the re-design results (i.e. it isolates the effects of different bandwidth sizes). Thus for the univariate re-designs, we specify a bandwidth of 5% for both the GWSD and GWMAD calibrations. For the bivariate re-designs, we specify a bandwidth of 41% for basic and robust GWCOR calibrations. For the multivariate re-designs, we specify a bandwidth of 52% for basic and robust GWPCA calibrations. These bandwidths are averages of those found within each group. Observe the benefits of conducting an extensive bandwidth study, as our final selections are more assured. The bandwidth function in any GW method should always be thoroughly investigated, and can be considered analogous to a thorough investigation of the variogram in geostatistics; where both investigations aim to identify spatial structure in some way (see Cressie 1989).

5.2 Local analyses

Maps of the demand data are now scrutinised for the univariate, bivariate and multivariate re-designs. Figs. 1c-d present the GWSD and GWMAD maps for pH, where high levels of pH variation tend to cluster at sites in northern England. This can be attributed to changes in the sources of acidification (natural and anthropogenic), as well as changes in land use and geology; all operating across small distances. Thus for the corresponding re-designs, these particular regions will be preferentially sampled. Note that the average GWSD value is slightly higher than the average GWMAD value, thus a GWMAD performs as expected in reducing variability at most locations (as a likely consequence of outlying pH data). However, minimum and maximum GWSD values are respectively, larger and smaller than that found using a GWMAD.

Figs. 3a-b present the robust and basic, bivariate GWPCA PTV maps for aluminium and TOC (which can only be for the first component, PC1). Figs. 3c-d present the corresponding robust and basic GWCOR maps; data that will be used for the bivariate re-designs. All calibrations use the same bandwidth of 41%. For bivariate GWPCA to act as a reasonable surrogate bandwidth selection method for GWCOR, we expect the GWPCA PTV data to roughly correlate with the $|\text{GWCOR}|$ data. In this respect, basic correlations between these outputs are promising at 0.91 and 0.52, for basic and robust cases, respectively. These correlations are also useful, in that they provide some clarity to our second and third postulations of Sect. 3.4. From the GWCOR maps, weak GWCOR values tend to cluster in all regions of England and Wales, thus for the corresponding re-designs, the same regions will be preferentially sampled. The strong positive GWCOR values observed in Scotland can be attributed to large areas of coniferous woodlands promoting aluminium mobilisation when freshwaters are acidified together with naturally occurring humic waters with high TOC. Some small regional differences can be observed between the basic and robust GWCOR data, most notably in south Wales, where weak negative correlations change to weak positive correlations.

For the basic and robust multivariate re-designs, we show the spatial distribution of the GWPCA PTV data for the first (PC1) and the first two components combined (PC1-2) in Fig. 4. In all four maps, the spatial patterns in the PTV data are broadly similar, with the highest PTV values generally located in England and Wales, whilst the lowest PTV values are located in north-west Scotland. For much of Scotland, the lower PTV data indicates a regional reduction in variable collinearity, and thus these regions will be preferentially sampled in the corresponding re-designs. Scotland, as a whole, appears to have the most spatially-diverse water chemistry data structures. The banded south-west to north-east patterns broadly follow the distribution of major soil-types and the under-lying geology (for example, see the soil transacts maps at The Macaulay Land Use Research Institute at http://www.macaulay.ac.uk/tipss/scotst1.htm, last accessed 13/01/14). As would be expected, variability in the PC1 PTV data is higher than that found in the PC1-2 PTV data. PTV data variation simply declines towards zero, as more components are combined. Thus in the context of our re-design study, PTV data for the first or the first few components combined, should only be used as demand data for the L-A algorithm. Although the spatial trends in the basic and robust GWPCA PTV data are broadly similar (at least at the large scale), the robust data are consistently larger. Furthermore, for robust GWPCA, its PTV data is always larger in the local case, than that found in the global case (with a robust PCA). At the local scale, the largest relative differences between the basic and robust GWPCA PTV data (for PC1 and for PC1-2) actually occur on the north-western coast and nearby isles of Scotland. Differences between basic and robust GWPCA PTV data can be taken to indicate the existence of multivariate outliers, some of which are likely to be locally-outlying.

5.3 Network re-designs

The L-A algorithm is first run with the eight demand data sets found above; in each case, targeting 25 of 533 sites for a second sampling campaign. Two maps for the resultant univariate re-designs are given in Figs. 5a-b, for basic and robust cases. Similarly, two maps for the resultant bivariate re-designs are given in Figs. 5c-d. Four maps are presented in Fig. 6 for the resultant multivariate re-designs, where basic and robust GWPCA PTV data for PC1 and for PC1-2 are used as the demand data sets. A ninth run of the L-A algorithm uses a demand data set of 533 equal values, resulting in a benchmark re-design for 25 sites (Fig. 7c). This benchmark re-design solely accounts for the particular spatial configuration of the study data; a configuration that directly influences the characteristics of all of the previous eight demand data sets (see for example, the preferential/clustered sampling studies of Olea 2007; Diggle et al. 2010; Gelfand et al. 2012). In this benchmark re-design, the consequences of over- and under-sampling on the re-design are isolated, such as: (i) the data void in central England, a region where freshwater acidification damage was considered unlikely; or (ii) the uneven sampling across 10km or 20km grid squares.

Thus each of the eight, GW method-based re-designs must first and foremost be compared to the benchmark re-design. Only those re-designs that deviate from the benchmark re-design are of interest and warrant further scrutiny. On viewing the eight re-designs, all clearly deviate from the benchmark re-design, where the smallest (or most subtle) deviations are found in the two bivariate cases.

All eight re-designs appear reasonable, considering the spatial distribution of the demand data (Figs. 1c-d, 3c-d and 4) and the spatial configuration of the initial network of sites. As expected, sites are preferentially located in northern England in the univariate case, coinciding with high levels of pH variation. As expected, sites are preferentially located in England and Wales in the bivariate case, coinciding with weak aluminium and TOC correlations. Here, in comparison to the benchmark re-design, there is one less site in Scotland (view sites on the west coast), whilst there is one more site in England (view sites in the south). Also as expected, sites are preferentially located in Scotland in the multivariate case, where thirteen to fourteen to sites are now located in Scotland compared to eleven in the benchmark re-design. As would be expected from the observed differences in the GWSD and GWMAD data; the basic and robust GWCOR data; and the four different GWPCA PTV data sets, the respective re-designs depend on which GWSD/GWMAD; GWCOR; and GWPCA specification is preferred. For example, on viewing the robust multivariate re-design (with PTV PC1-2 data in Fig. 6d), notable differences with the corresponding basic re-design (Fig. 6c) are found in: (a) Scotland, where one extra site is located; (b) Wales, where one less site is located; and (c) southern England, where the same four sites are chosen but *demand* is reduced (i.e. smaller pink circles).

Finally, two further runs of the L-A algorithm are conducted with weighted demand data to provide examples of weighted (or biased) re-designs. There are various options to weight our demand data sets, such as those which: (1) account for the accessibility of a freshwater site, in order to reduce sampling costs at remote locations or (2) reflect the buffering capacity of a freshwater site to acidification. Population data may act as simple surrogate weighting data set for the former option, whilst freshwater acidification critical load data can be used for the latter option. As the critical load data are a constituent part of the study data set, and available at all 533 freshwater sites, we use this data to provide examples of weighted re-designs. In particular, we weight our robust GWPCA PTV data sets for PC1 and for PC1-2, by $1/\exp(\mathrm{SSWC})$, where SSWC are critical load data calculated via the steady-state water chemistry model (Henriksen et al. 1992). The (non-geographical) weighting reflects the fact that a critical load of above 5 keq $H^{+}ha^{-1}year^{-1}$ is unlikely to be exceeded by its corresponding acid deposition value (and thus irreversible damage is unlikely to occur). These weighted, multivariate re-designs are given in Figs. 7a-b, along with a map of the (positively skewed) SSWC data (Fig. 7d). Both re-designs display a strong clustering of sites in northern Scotland, a region of low critical loads coinciding with low PTV data. Re-designs represent a targeted sampling campaign in regions of most concern to ecological damage via acidification, whilst ensuring that regions, less susceptible to damage, are not unduly under-sampled.

6 Conclusions

In summary, a location-allocation (L-A) algorithm calibrated with outputs from a geographically weighted (GW) method provides a means to optimally design a second sampling campaign for univariate, bivariate or multivariate spatial data sets. The network re-design procedure preserves spatial and attribute information; and is considered an extension of the univariate-only procedure of Kanaroglou et al. (2005) used to optimally locate air pollution monitors. In this respect, it is applicable to any pollution study, be it water (as in this study), soil or air pollution, where a high number of contaminants are routinely measured. Furthermore, the re-designs can be tailored (or weighted) so that targeted sampling is possible at sites where urgent management actions may be required to prevent environmental degradation.

As the collection of large environmental data sets continues (e.g. as required for the European Union Water Framework Directive), the availability of such re-design procedures are becoming increasingly valuable. GW method-based re-design procedures are relatively simple in comparison to geostatistical alternatives, where a useful comparison of both approaches is left for future work, possibly with simulated data. Future work is also expected concerning the adaptation of our re-design procedures with: (i) alternative L-A algorithms; (ii) different distance metrics, say for soil geochemistry re-designs in urban areas (e.g. Glennon et al. 2014) and (iii) GW variogram-type measures (Harris et al. 2010) extended to multivariate forms.

Acknowledgements

For Harris and Charlton, research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by the Science Foundation Ireland under the National Development Plan.

References

Baume OP, Gebhardt A, Gebhardt C, Heuvelink GBM, Pilz J (2011) Network optimization algorithms and scenarios in the context of automatic mapping. Comput Geosci 37:289-294

Brunsdon C, Fotheringham AS, Charlton M (1998) Geographically weighted regression: modelling spatial non-stationarity. J Roy Stat Soc D-Sta 47:431-443

Brunsdon C, Fotheringham AS, Charlton M (2002) Geographically weighted summary statistics - a framework for localised exploratory data analysis. Comput Environ Urban 26:501-524

Brus DJ, de Gruijter J (1997) Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil. Geoderma 80:1-59

Brus DJ, Heuvelink GBM (2007) Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138:86-95

Caselton WF, Zidek JV (1984) Optimal monitoring networks. Stat Probabil Lett 2:223-227

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

CLAG Freshwaters (1995) Critical Loads of Acid Deposition for United Kingdom Freshwaters. Critical Loads Advisory Group, Sub-report on Freshwaters, ITE, Penicuik

Cressie NA (1989) The many faces of spatial prediction. In M. Armstrong (Eds.), Geostatistics - Vol 1 (pp. 163-176). Dordrecht: Kluwer

de Gruijter J, Brus D, Bierkens M, Knotters M (2006) Sampling for Natural Resource Monitoring. Springer, New York

Delmelle EM, Goovaerts P (2009) Second-phase sampling designs for non-stationary spatial variables. Geoderma 153:205-216

Diggle PJ, Menezes R, Su T (2010) Geostatistical inference under preferential sampling. J Roy Stat Soc C-Sta 59:191–232

Farber S, Páez A (2007) A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. J Geogr Syst 9:371-396

Filzmoser P, Todorov V (2012) Robust tools for the imperfect world. Inform Sciences DOI 10.1016/j.ins.2012.10.017

Fotheringham AS, Brunsdon C, Charlton M (2002) Geographically Weighted Regression - the analysis of spatially varying relationships. Wiley, Chichester

Gelfand AE, Sahu SK, Holland DM (2012) On the effect of preferential sampling in spatial prediction. Environmetrics, 23:565–578

Glennon M, Harris P, Finne T, Scanlon R, O'Connor P (2014) Geochemical baseline for heavy metals in topsoils in Dublin, Ireland: spatial correlation with historic industry and implications for human health. Environ Geochem Hlth DOI 10.1007/s10653-013-9561-8

Griffith DA (2005) Effective geographic sample size in the presence of spatial autocorrelation. Ann Assoc Am Geogr 95:740-760

Haas TC (1992) Redesigning continental-scale monitoring networks. Atmos Environ 26A:3323-3333

Haas TC (2002) New systems for modelling, estimating, and predicting a multivariate spatio-temporal process. Environmetrics 13:311-332

Hampel FR (1974) The influence curve and its role in robust estimation. J Am Stat Assoc 69:383-393

Harris P, Charlton M, Fotheringham AS (2010) Moving window kriging with geographically weighted variograms. Stoch Environ Res Risk Assess 24:1193-1209

Harris P, Brunsdon C, Charlton M (2011) Geographically weighted principal components analysis. Int J Geogr Inf Sci 25:1717-1736

Harris P, Charlton M, Brunsdon C (2012) Geographically weighted (GW) models: advances in modelling spatial heterogeneity. geoENV 2012, Valencia, Spain

Harris P, Brunsdon C, Charlton M, Juggins S, Clarke A (2014) Multivariate spatial outlier detection using robust geographically weighted methods. Math Geosci DOI 10.1007/s11004-013-9491-0

Henriksen A, Kämäri J, Posch M, Wilander A (1992) Critical loads of acidity: Nordic surface waters. Ambio 21:356-363

Holmes JF, Williams FB, Brown LA (1972) Faculty location under a maximum travel restriction: An example using day care facilities. Geogr Anal 4:258-266

Hornung M, Bull KR, Cresser M, Ullyett J, Hall JR, Langan S, Loveland PJ, Wilson

MJ (1995) The sensitivity of surface waters of Great Britain to acidification predicted from catchment characteristics. Environ Pollut 87:207-214

Jolliffe IT (2002) Principal Components Analysis. 2nd edition. Springer-Verlag, New York

Journel AG (1986) Geostatistics: models and tools for the earth sciences. Math Geol 18:119-140

Kanaroglou PS, Jerrett M, Morrison J, Beckerman B, Arain MA, Gilbert NL, Brook JR (2005) Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach. Atmos Environ 39:2399-2409

Kreiser AM, Patrick ST, Battarbee RW (1993) Critical loads for UK freshwaters - introduction, sampling strategy and use of maps. In: Hornung, M., Skeffington, R.A. (Eds.) Critical loads: Concepts and Applications. ITE symposium No.28, HMSO, London, pp. 94-98

Le ND, Zidek JV (1992) Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. J Multivariate Anal 43:351-374

Le ND, Zidek JV (2006) Statistical Analysis of Environmental Space-Time Processes. Springer, New York

Lu B, Harris P, Gollini I, Charlton M, Brunsdon C (2013) GWmodel: an R package for exploring spatial heterogeneity. GISRUK 2013, Liverpool, UK

Marchant BP, Newman S, Corstanje R, Reddy KR, Osborne TZ, Lark RM (2009) Spatial monitoring of a non-stationary soil property: phosphorus in a Florida water conservation area. Eur Jour Soil Sci 60:759-769

Maronna R, Martin D, Yohai V (2006) Robust Statistics: Theory and Methods. Wiley, Toronto

Martin RJ (2001) Comparing and contrasting some environmental and experimental design problems. Environmetrics 12:273-287

McBratney AB, Webster R, Burgess TM (1981) The design of optimal sampling schemes for local estimation and mapping of regionalized variables. I. Theory and method. Comput Geosci 7:331-334

Müller WG (2005) A comparison of spatial design methods for correlated observations. Environmetrics 16:495-505

Müller WG (2007) Collecting Spatial Data. Springer, Heidelberg

Müller WG, Zimmerman DL (1999) Optimal designs for variogram estimation. Environmetrics 10:23-37

Olea RA (2007) Declustering of clustered preferential sampling for histogram and semivariogram inference. Math Geol 39:453-467

Pretty JN, Mason CF, Nedwell DB, Hine RE, Leaf S, Dils R (2003) Environmental Costs of Freshwater Eutrophication in England and Wales. Environ Sci Technol 37:201–208

ReVelle CS, Eiselt HA (2005) Location analysis: A synthesis and survey. Eur J Oper Res 165:1-19

Ritter K (1996) Asymptotic optimality of regular sequence designs. Ann Stat 24:2081-2096

Rosing K, Hillsman E, Rosing-Vogelaar H (1979) The robustness of two common heuristics for the P-median problem. Environ Plann A 11:373-380

Rousseeuw PJ, Debruyne M, Engelen S, Hubert M (2006) Robustness and outlier detection in chemometrics. Crit Rev Anal Chem 36: 221-242

Royle JA, Nychaka D (1998) An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Comput Geosci 24:479-488

Teitz MB, Bart P (1968) Heuristic methods for estimating the generalized vertex median of a weighted graph. Oper Res 16:955-961

van Groenigen J-W, Stein A (1998) Constrained optimisation of spatial sampling using continuous simulated annealing. J Environ Qual 27:1078-1086

van Groenigen J-W, Siderius W, Stein A (1999) Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87:239-259

van Groenigen J-W, Pieters G, Stein A (2000) Optimizing spatial sampling for multivariate contamination in urban areas. Environmetrics 11:227-244

Varmuza K, Filzmoser P (2009) Introduction to Multivariate Statistical Analysis in Chemometrics. CRC press

Vašát R, Heuvelink GBM, Borůvka L (2010) Sampling design optimization for multivariate soil mapping. Geoderma 155:147-153

Wang J, Haining R, Cao Z (2010) Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. Int J Geogr Inf Sci 24:532-543

Wang J, Stein A, Gao B, Ge Y (2012) A review of spatial sampling. Spatial Statistics 2:1-14

Warrick AW, Myers DE (1987) Optimisation of sampling locations for variogram calculations. Water Resour Res 23:496-500

Webster R, Welham SJ, Potts JM, Oilver MA (2006) Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood. Comput Geosci 32:1320-1333

Xia G, Miranda M, Gelfand AE (2006) Approximately optimal spatial design approaches for environmental health data. Environmetrics 17:363-385

Zhu Z, Stein ML (2005) Spatial sampling design for parameter estimation of the covariance function. J Stat Plann Inference 134:583-603

Zidek JV, Sun W, Le ND (2000) Designing and integrating composite networks for monitoring multivariate Gaussian pollution fields. J Roy Stat Soc C-Sta 49:63-79

Zimmerman DL (2006) Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. Environmetrics 17:635-652

Tables

Table 1 List of abbreviations and definitions.

Abbrev.	Definition	Abbrev.	Definition
GW	Geographically Weighted	GOF	Goodness Of Fit
L-A	Location-Allocation	RMSPE	Root Mean Squared Prediction Error
GWSS	GW Summary Statistics	MAPE	Mean Absolute Prediction Error
PCA	Principal Components Analysis	Alk.T	Alkalinity (μeq L^{-1}) (transformed)
GWPCA	GW Principal Components Analysis	Cond.T	Conductivity (μS cm^{-1}) (transformed)
NO_2	Nitrogen dioxide	NO3.T	Nitrate or NO_3^- (μeq L^{-1}) (transformed)
GWR	GW Regression	SO4.T	Sulphate or SO_4^{2+} (μeq L^{-1}) (transf.)
SD	Standard Deviation	PO4.T	Phosphate or PO_4 (μeq L^{-1}) (transf.)
MAD	Median Absolute Deviation	AL.TM.T	Total Monomeric Aluminium (μg L^{-1}) (transformed)
MCD	Minimum Covariance Determinant	TOC.T	Total Organic Carbon (mg L^{-1}) (transf.)
GWSD	GW Standard Deviation	TOC	Total Organic Carbon
GWMAD	GW Median Absolute Deviation	PC1	First principal component
GWCOR	GW Correlation	PC1-2	First two principal components combined
PTV	Percentage of the Total Variance	SSWC	Steady-State Water Chemistry

Table 2 Basic and robust global estimates of variability.

	pH	Alk.T	Cond.T	NO3.T	SO4.T	PO4.T	AL.TM.T	TOC.T
SD	1.07	2.23	0.90	6.34	1.11	6.41	0.80	0.42
MAD	1.08	2.38	0.92	5.49	1.07	5.45	0.62	0.44

Table 3 Basic and robust global correlation coefficients.

	pH	**Alk.T**	**Cond.T**	**NO3.T**	**SO4.T**	**PO4.T**	**AL.TM.T**	**TOC.T**
BASIC:								
pH	1	0.92	0.58	0.15	0.47	0.19	-0.63	0.10
Alk.T		1	0.75	0.21	0.63	0.33	-0.57	0.26
Cond.T			1	0.30	0.87	0.35	-0.33	0.41
NO3.T				1	0.39	0.18	0.00	-0.06
SO4.T					1	0.34	-0.21	0.34
PO4.T						1	-0.03	0.44
AL.TM.T							1	0.11
TOC.T								1
ROBUST:								
pH	1	0.93	0.58	0.21	0.52	0.17	-0.66	0.04
Alk.T		1	0.74	0.29	0.67	0.30	-0.59	0.19
Cond.T			1	0.37	0.87	0.31	-0.33	0.34
NO3.T				1	0.47	0.21	-0.03	-0.03
SO4.T					1	0.33	-0.22	0.30
PO4.T						1	0.03	0.45
AL.TM.T							1	0.29
TOC.T								1

Table 4 Eigenvalues, PTV (%) and cumulative PTV (%) from basic and robust PCA.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
BASIC:								
Eigenvalues:	3.78	1.53	1.08	0.69	0.40	0.36	0.11	0.04
PTV:	47.3	19.1	13.5	8.7	5.0	4.5	1.4	0.6
Cumulative PTV:	47.3	66.4	79.9	88.6	93.5	98.1	99.4	100.0
ROBUST:								
Eigenvalues:	3.91	1.67	1.12	0.73	0.41	0.23	0.11	0.04
PTV:	47.5	20.4	13.6	8.9	5.0	2.8	1.3	0.5
Cumulative PTV:	47.5	68.0	81.6	90.4	95.4	98.2	99.5	100.0

Table 5 Basic and robust bandwidths (%) for GW methods & associated re-designs.

	Basic	**Robust**
UNIVARIATE RE-DESIGN:		
GW mean as surrogate for GWSD (pH)	3	MNR
GW median as surrogate for GWMAD (pH)	7	7
BIVARIATE RE-DESIGN:		
Basic bivariate GWPCA as surrogate for GWCOR (AL.TM.T and TOC.T)	33*	45*
Robust bivariate GWPCA as surrogate for GWCOR (AL.TM.T and TOC.T)	41*	45*
MULTIVARIATE RE-DESIGN:		
Basic multivariate GWPCA with $\boldsymbol{q} = 4$	48	56
Robust multivariate GWPCA with $\boldsymbol{q} = 4$	MNR	MNR

MNR = Minimum Not Reached;

* Judged optimum from interpretation of bandwidth function.

Figures

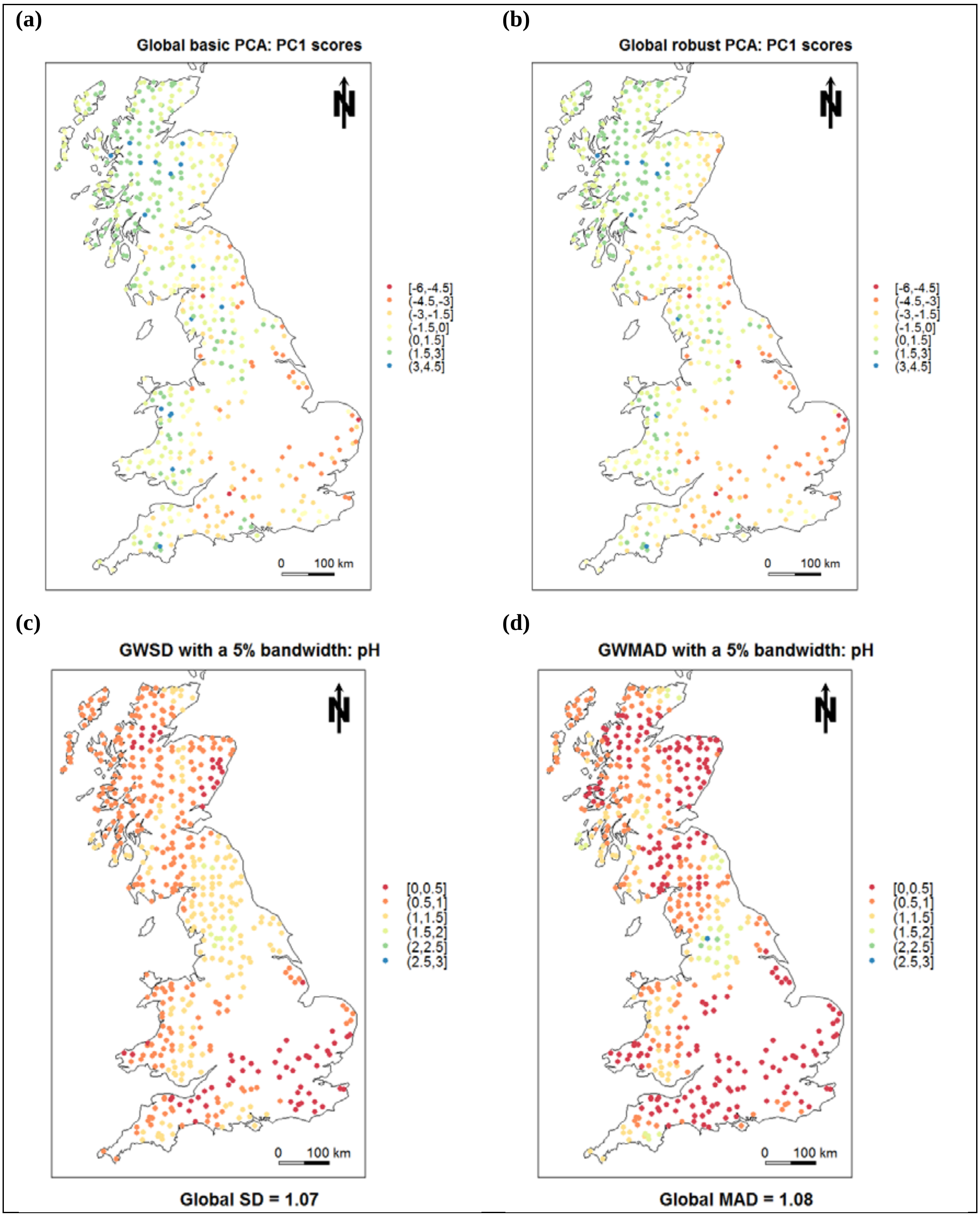

Fig. 1 PCA scores for the first component (PC1): **(a)** basic and **(b)** robust. Local variability maps for univariate re-designs with pH data: **(c)** GWSD and **(d)** GWMAD.

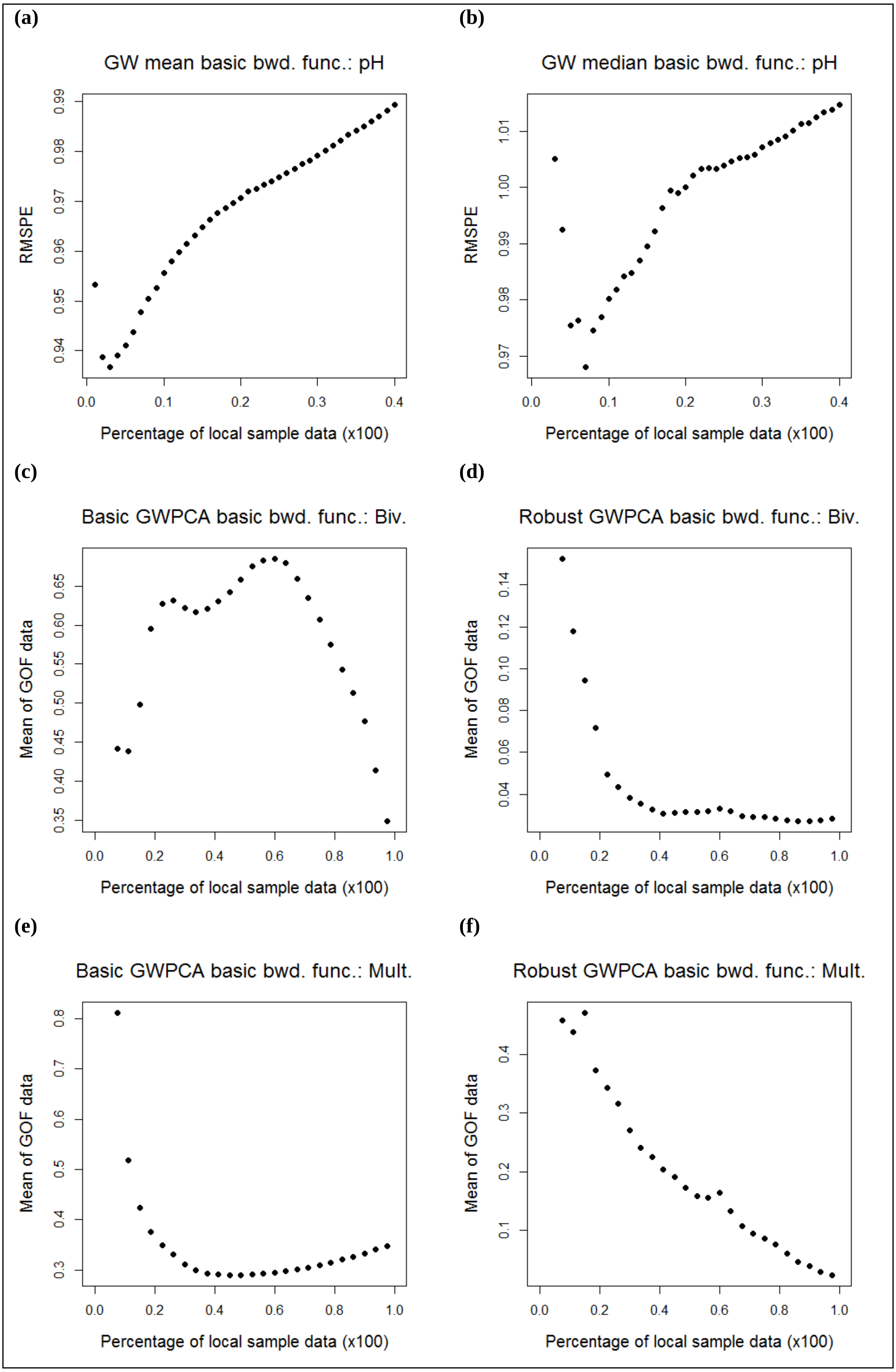

Fig. 2 Basic bandwidth functions for respective, basic and robust: **(a-b)** univariate, **(c-d)** bivariate, and **(e-f)** multivariate re-designs.

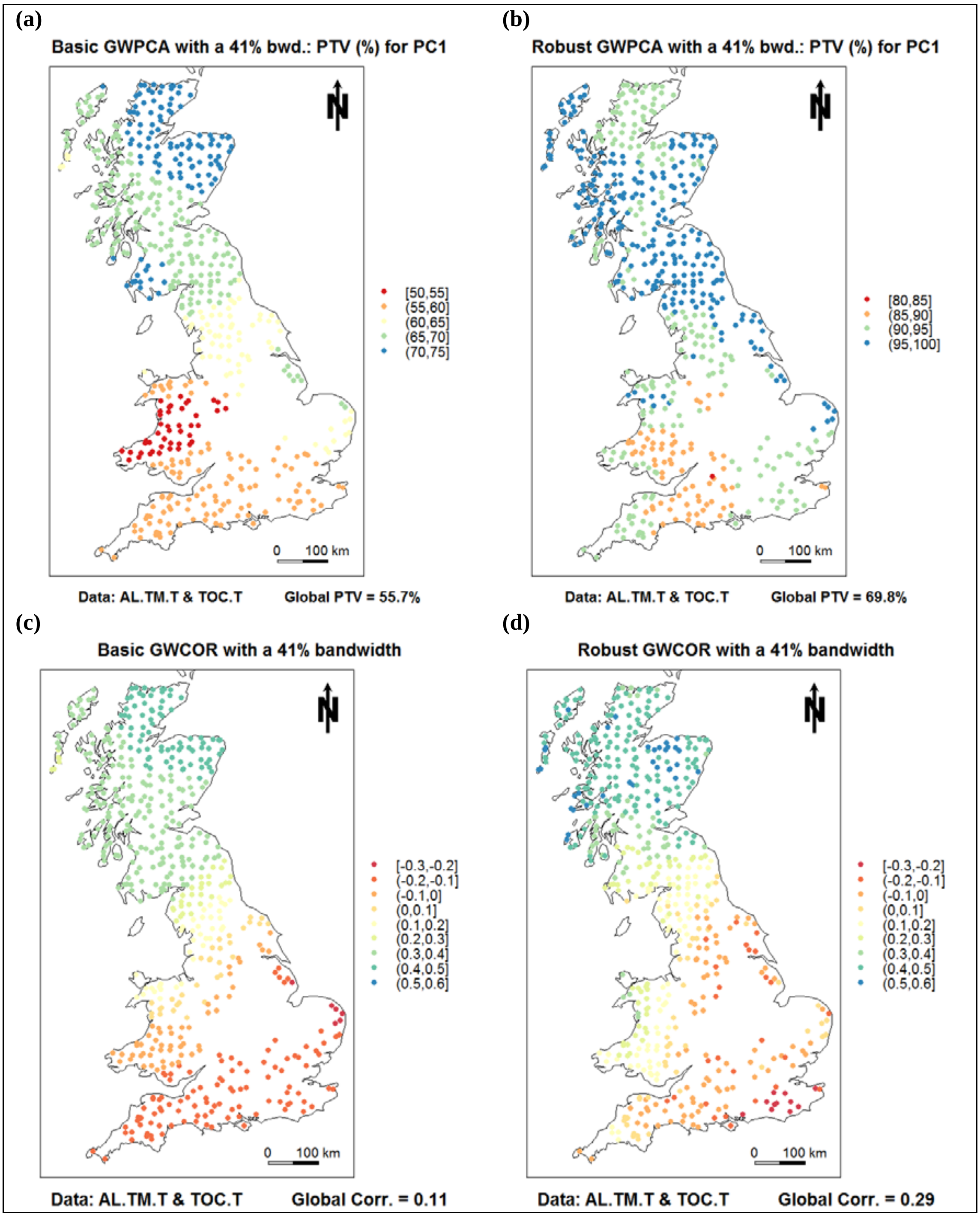

Fig. 3 Maps for bivariate re-designs with AL.TM.T and TOC.T: **(a)** basic GWPCA PTV data, **(b)** robust GWPCA PTV data, **(c)** basic GWCOR, and **(d)** robust GWCOR.

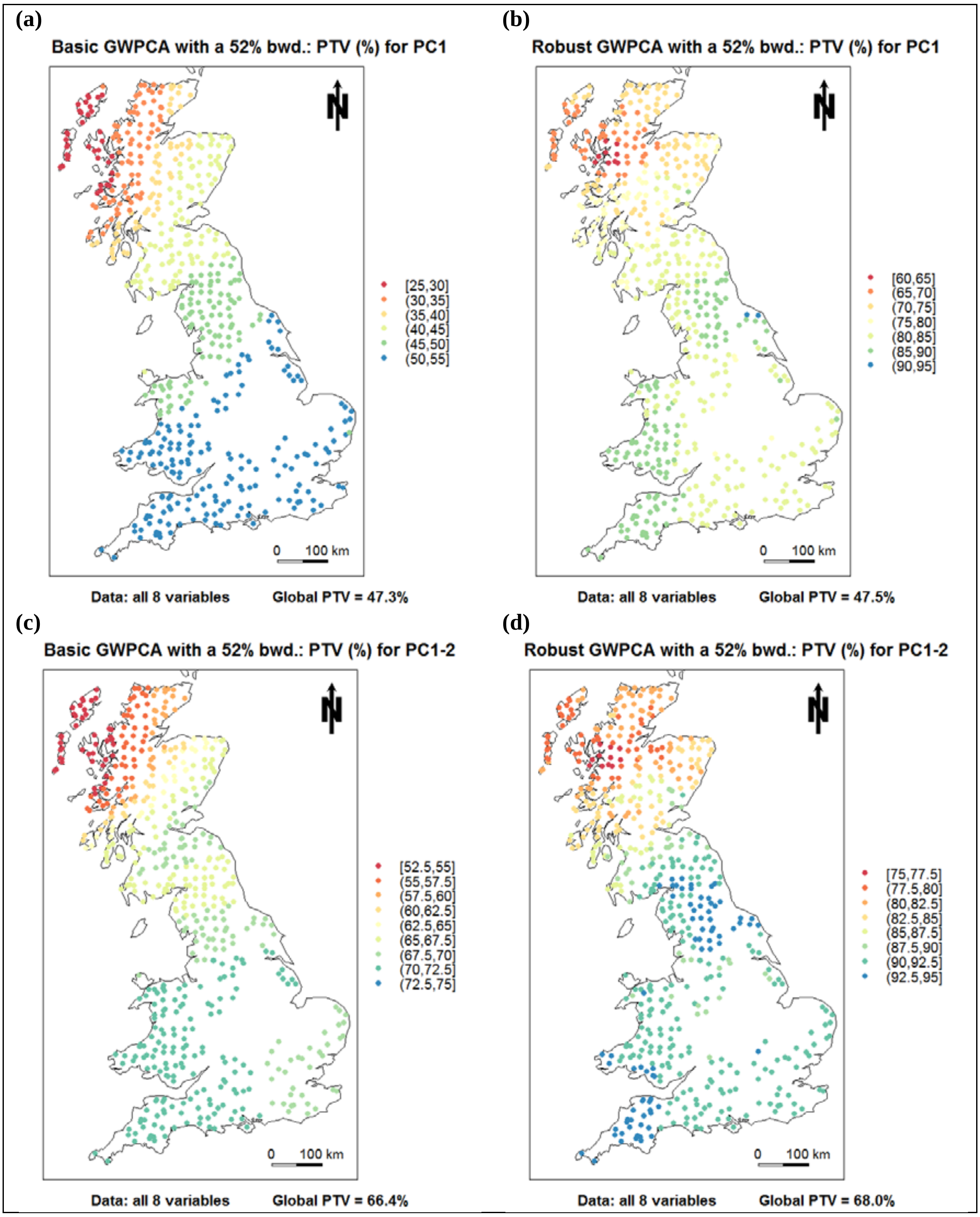

Fig. 4 GWPCA PTV maps for multivariate re-designs: **(a)** basic with PC1 only, **(b)** robust with PC1 only, **(c)** basic with PC1 plus PC2, and **(d)** robust with PC1 plus PC2.

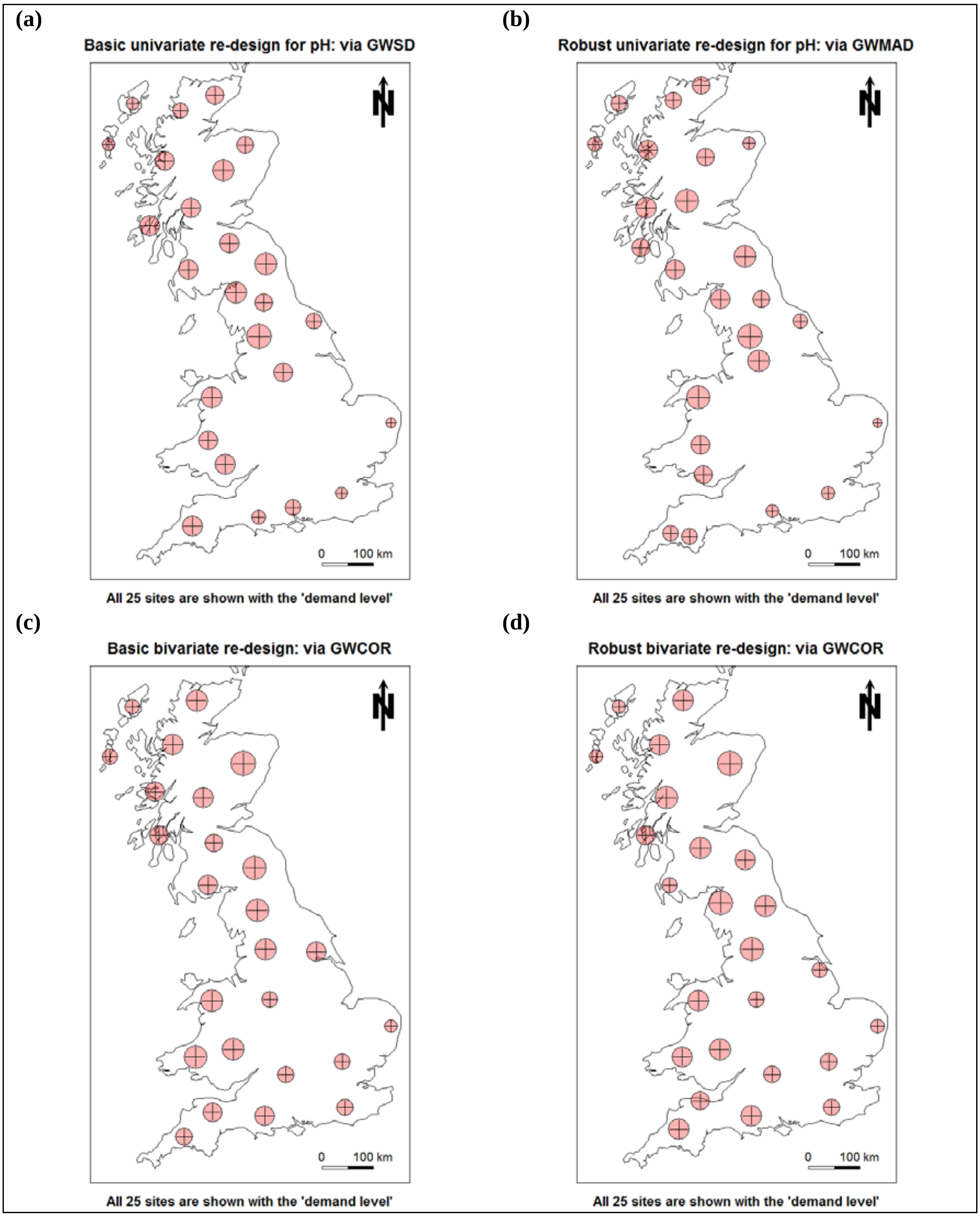
(a)
Basic univariate re-design for pH: via GWSD
0
100 km
All 25 sites are shown with the 'demand level'
(b)
Robust univariate re-design for pH: via GWMAD
0
100 km
All 25 sites are shown with the 'demand level'
(c)
Basic bivariate re-design: via GWCOR
0
100 km
All 25 sites are shown with the 'demand level'
(d)
Robust bivariate re-design: via GWCOR
0
100 km
All 25 sites are shown with the 'demand level'

Fig. 5 Re-designs for: **(a)** basic univariate, **(b)** robust univariate, **(c)** basic bivariate, and **(d)** robust bivariate. Demand level is reflected by the size of the pink circle at a chosen site.

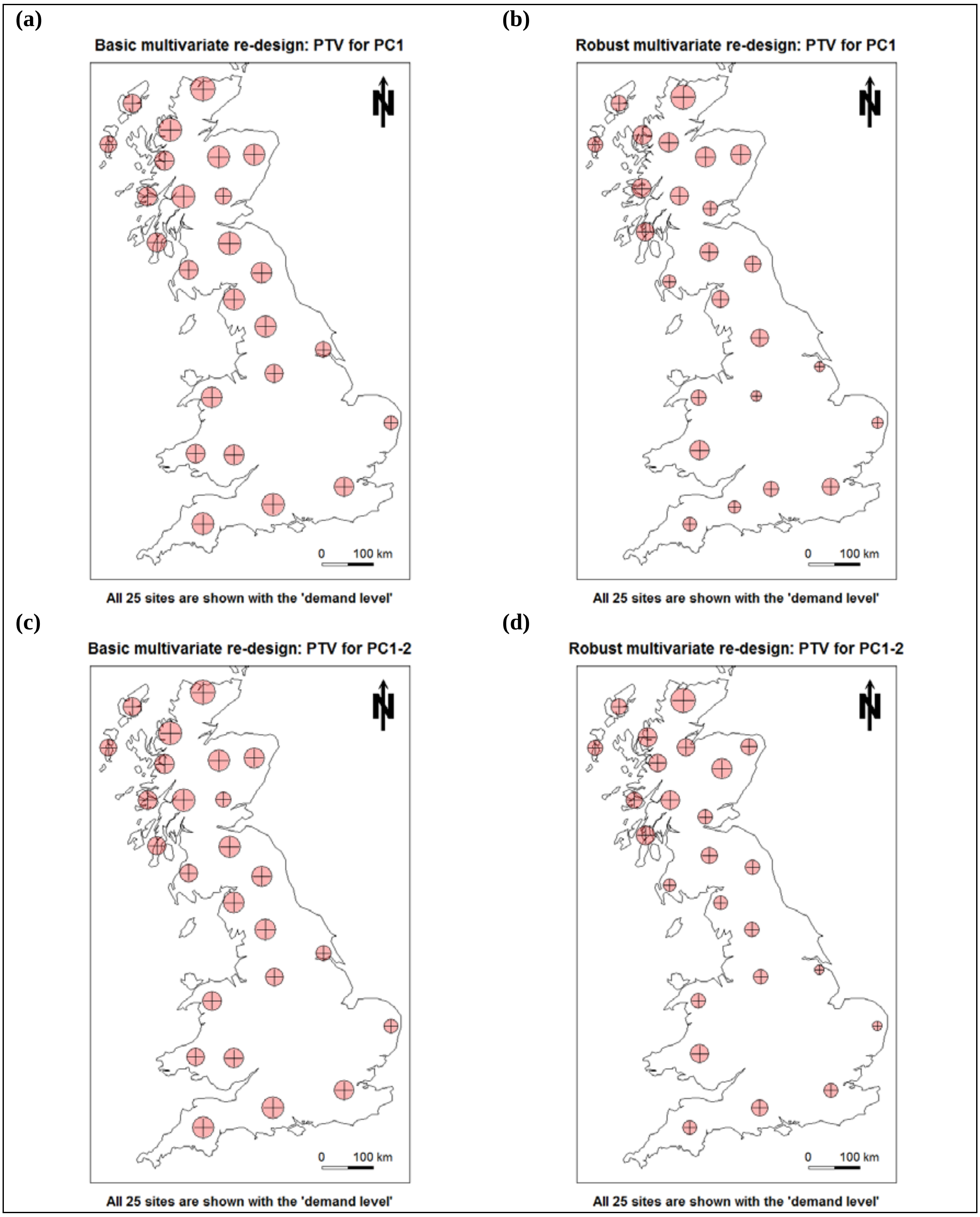
(a)
Basic multivariate re-design: PTV for PC1
0
100 km
All 25 sites are shown with the 'demand level'
(b)
Robust multivariate re-design: PTV for PC1
0
100 km
All 25 sites are shown with the 'demand level'
(c)
Basic multivariate re-design: PTV for PC1-2
0
100 km
All 25 sites are shown with the 'demand level'
(d)
Robust multivariate re-design: PTV for PC1-2
0
100 km
All 25 sites are shown with the 'demand level'

Fig. 6 Multivariate re-designs for: **(a)** basic via PC1 data, **(b)** robust via PC1 data, **(c)** basic via PC1 plus PC2 data, and **(d)** robust via PC1 plus PC2 data. Demand level is reflected by the size of the pink circle at a chosen site.

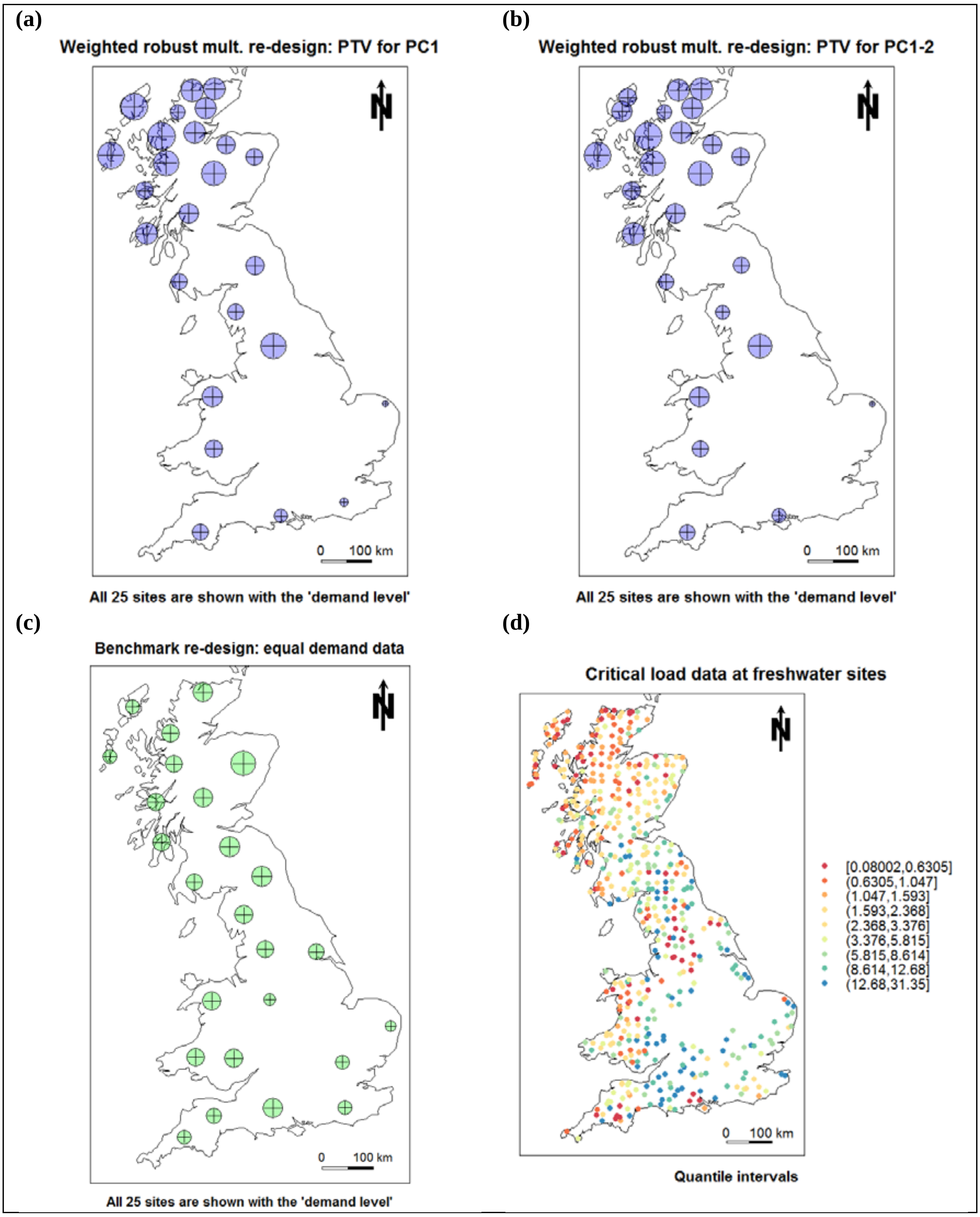
(a)
Weighted robust mult. re-design: PTV for PC1
0 100 km
All 25 sites are shown with the 'demand level'
(b)
Weighted robust mult. re-design: PTV for PC1-2
0 100 km
All 25 sites are shown with the 'demand level'
(c)
Benchmark re-design: equal demand data
0 100 km
All 25 sites are shown with the 'demand level'
(d)
Critical load data at freshwater sites
0 100 km
[0.08002,0.6305]
(0.6305,1.047]
(1.047,1.593]
(1.593,2.368]
(2.368,3.376]
(3.376,5.815]
(5.815,8.614]
(8.614,12.68]
(12.68,31.35]
Quantile intervals

Fig. 7 Weighted re-designs (multivariate and robust): **(a)** via PC1 data and **(b)** via PC1 plus PC2 data. Benchmark re-design with equal demand data given in **(c).** Demand level is reflected by the size of the blue or green circle at a chosen site. Freshwater acidification critical load map given in **(d)**.