Math Geosci (2014) 46:1–31
DOI 10.1007/s11004-013-9491-0

Multivariate Spatial Outlier Detection Using Robust Geographically Weighted Methods

Paul Harris · Chris Brunsdon · Martin Charlton · Steve Juggins · Annemarie Clarke

Received: 4 February 2013 / Accepted: 6 September 2013 / Published online: 26 September 2013

Abstract Outlier detection is often a key task in a statistical analysis and helps guard against poor decision-making based on results that have been influenced by anomalous observations. For multivariate data sets, large Mahalanobis distances in raw data space or large Mahalanobis distances in principal components analysis, transformed data space, are routinely used to detect outliers. Detection in principal components analysis space can also utilise goodness of fit distances. For spatial applications, however, these global forms can only detect outliers in a non-spatial manner. This can result in false positive detections, such as when an observation's spatial neighbours are similar, or false negative detections such as when its spatial neighbours are dissimilar. To avoid mis-classifications, we demonstrate that a local adaptation of various global methods can be used to detect multivariate spatial outliers. In particular, we account for local spatial effects via the use of geographically weighted data with either Mahalanobis distances or principal components analysis. Detection performance is assessed using simulated data as well as freshwater chemistry data collected over all of Great Britain. Results clearly show value in both geographically weighted methods to outlier detection.

P. Harris (✉) · M. Charlton
National Centre for Geocomputation, National University of Ireland Maynooth, Iontas Building,
Maynooth, Co. Kildare, Ireland
e-mail: Paul.Harris@nuim.ie

C. Brunsdon
Geography and Planning, University of Liverpool, Liverpool, England, UK

S. Juggins
School of Geography, Politics and Sociology, University of Newcastle, Newcastle upon Tyne,
England, UK

A. Clarke
APEM Ltd, Llantrisant, Wales, UK

Keywords Non-stationarity · Mahalanobis distance · Principal components analysis · Co-kriging cross-validation · Freshwater acidification · Anomaly detection

1 Introduction

Outlier identification is often a key task in a statistical analysis and helps guard against poor decision-making based on results that have been adversely or beneficially influenced by anomalous observations. Anomalous or exceptional data values may represent: (a) data recording or measurement errors or (b) true data values that are atypical. In the former case, outlier detection serves as a useful data cleaning or screening exercise, whilst in the latter case, outlier detection can uncover interesting or unusual properties in the data that may have gone un-noticed. Further, when numerous (true data) outliers are detected, this may provide evidence of more than one under-lying population. Populations may operate locally or on different observational scales.

In geographical settings, outliers can have any combination of non-spatial, relationship, spatial, or temporal characteristics, as depicted in Fig. 1. Non-spatial (univariate) outliers simply reflect the occurrence of an observation lying in one of the tails of its sample distribution. A simple boxplot analysis can be used to detect these outliers (e.g. Hubert and Vandervieren 2008). In the bivariate case, a relationship outlier occurs when a data pair at a given observation point is unusual in relation to the behaviour of all other data pairs and bagplots can be used as a method of detection (Rousseeuw et al. 1999). Extension to the multivariate case is analogous and relates to when a vector of data at a given observation point is unusual with respect to all other observation data vectors. Here, Mahalanobis distances (MDs) to the centre of the multivariate data set can be calculated, where large MDs are associated with outliers (e.g. Filzmoser et al. 2005). Spatial (univariate) outliers arise when an observation is unusual with respect to its close spatial neighbours. Here, the intuitively anticipated positive local spatial autocorrelation is absent and a local spatial autocorrelation measure such as local Moran's I (Anselin 1995) can be used as a method of detection. Temporal (univariate) outliers are analogous to spatial outliers, but in one- not two-dimensions. Similarly, it follows that a temporal dependence measure can be used to detect these outliers (e.g. Ljung 1993). More recently, Sun and Genton (2011) used adjusted functional boxplots to detect outliers in space-time.

For each outlier-type, there are competing methods of detection. For multivariate outliers, various methods can be used, often depending on the dimensionality of the data (e.g. Rousseeuw et al. 2006; Daszykowski et al. 2007; Filzmoser and Todorov 2013). Similarly, for (univariate-only) spatial outliers, various methods are available (Krige and Magri 1982; Liu et al. 2001; Glatzer and Müller 2004; Kou et al. 2006; Chen et al. 2008). However, methods that combine both characteristics, accounting for the multivariate and spatial nature of the data are rare. The only known body of research in this area can be found in Lu et al. (2004); Chen et al. (2008), where robust local MDs are used to detect multivariate spatial outliers.

We similarly use robust local MDs for this purpose, but our study additionally investigates the use of robust local principal components analyses (PCA) as a method of

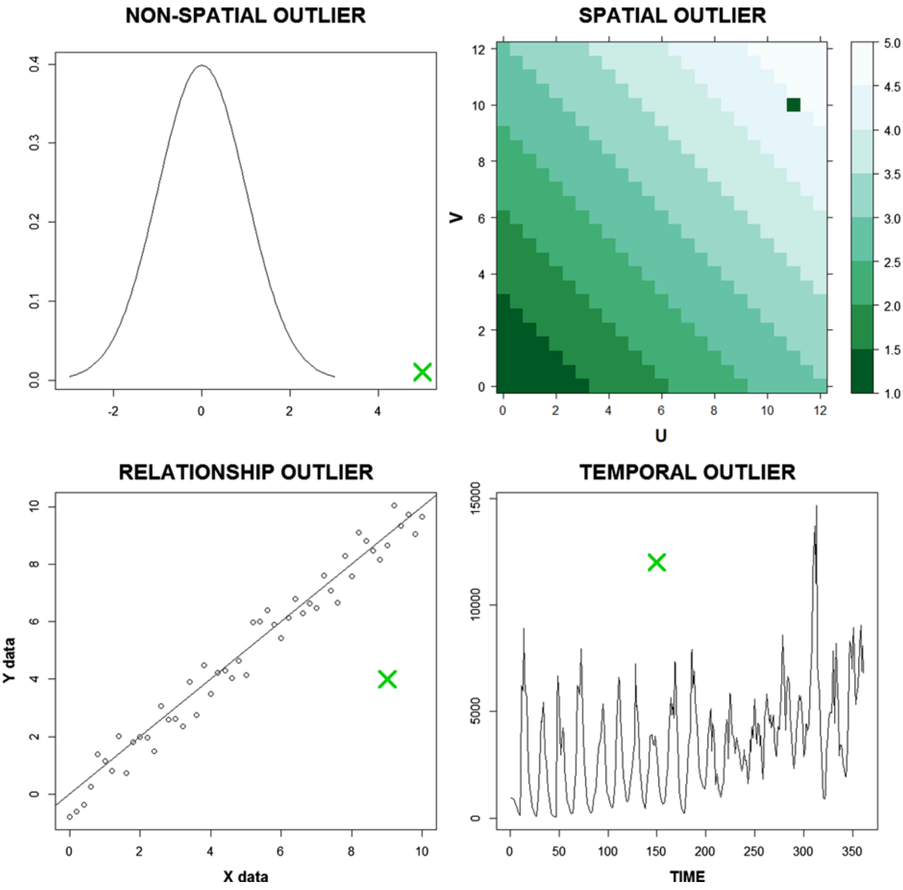

Fig. 1 Four types of outliers in geographical settings

detection. For our local methods, the data is geographically weighted (GW), as found in the GW methods of Fotheringham et al. (2002); resulting in novel robust GWMD and robust GWPCA detection methods. Observe that we specify robust methods. Detecting outliers using a basic (non-robust) method is not recommended as the outliers themselves can compromise a basic method's fit prior to its use as a method of detection; and as a result, the outliers are not detected. To avoid these effects, a robust method attempts to fit a model to the majority of the data; data that is least likely to include outliers, and in doing so, outliers are detected as those observations that deviate strongly from this robust fit. Thus, in the context of our study, a robust method is designed to work reliably with data contaminated by outliers, which in turn should ensure that key statistical assumptions (e.g. normality) are not violated (Rousseeuw et al. 2006; Filzmoser and Todorov 2013).

We evaluate our detection methods using: (i) simulated data and (ii) freshwater chemistry data for Great Britain (CLAG CLAG Freshwaters 1995). The design of the data simulation study is also considered an advance, where a geostatistical co-simulation algorithm is used to generate spatially co-dependent data prior to a contamination with outliers. The use of simulated data allows us to compare the detection performance of our robust GW methods with two benchmark methods: (A) a non-spatial MD method and (B) co-kriging (CoK) cross-validation that is spatial by design. Our study is structured as follows: (1) methodology, for details of our detection methods; (2) the design and results from the data simulation study; (3) results from the empirical study; and (4) conclusions. All GWMD and GWPCA functions were implemented in R (http://www.r-project.org) and will be made available in the *GW-model* R package in due course.

2 Methodology

For spatial applications, the use of a standard (global) MD- or PCA-based method to detect multivariate outliers can result in a false positive, when an observation's spatial neighbours are similar in value (i.e. the observation is not locally-outlying), or a false negative when its spatial neighbours are dissimilar in value (i.e. the observation is locally-outlying). To address these particular forms of misclassification, we describe local adaptations of global methods that can be used to detect multivariate spatial outliers. We do not envisage that a local method should replace its global counterpart, but instead, they should complement each other. The global method provides a broad, general sweep for outliers, whereas the local method provides a deep, more focused identification.

2.1 Robust Methods and Outlier Detection

Outliers are commonly identified by large residuals or deviations from some robust method's fit. Outliers cannot be so easily identified using a basic (non-robust) method, as the fit is so poor that the outliers are masked. Furthermore, on applying a basic method to data with outliers, data may be assigned as outlying when they are not, an effect known as swamping. Robust methods are possible for estimating location and scale/scatter in both univariate and multivariate cases (e.g. Daszykowski et al. 2007). For the multivariate case, robust estimates for the mean vector and the covariance matrix can be found concurrently, using for example, the minimum covariance determinant (MCD) estimator or the minimum volume ellipsoid estimator (Rousseeuw 1985; Filzmoser and Todorov 2013). For this study, we need to locally-specify such a robust estimator for our GWMD and GWPCA detection methods.

2.2 Geographically Weighted Methods

In this section, we present a brief overview of GW methods with respect to their main use in the exploration of spatial heterogeneity. These non-stationary methods suit situations when the data is poorly described by some universal or global model fit and where for some regions, a localised fit provides a better description. The approach uses a moving window weighting technique, where local models are calibrated at (sampled or un-sampled) target locations (i.e. the window's centre). For an individual model at some target location, we weight all neighbouring observations according to the properties of some distance-decay kernel function and then locally fit the model to this weighted data. Thus, the geographical weighting solely applies to the sample 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 function's bandwidth. Small bandwidths lead to more rapid spatial variation in the results, while large bandwidths yield results increasingly close to the universal model solution. When there exists some objective function (i.e. the model can be used as a spatial predictor), an optimal bandwidth can be found using cross-validation. Commonly, the local outputs or parameters of a given GW method are mapped to provide a useful exploratory tool, that can precede a more traditional (global) or sophisticated (local) statistical analysis.

Almost any statistical method can be extended to a GW form. The most popular is GW regression (Brunsdon et al. 1996; Fotheringham et al. 2002; Wheeler 2007), where local regressions are found at target locations. The resultant regression coefficients are then mapped to assess for spatial change in the relationships between the dependent and independent variables. Other GW methods include: GW summary statistics (Brunsdon et al. 2002; Fotheringham et al. 2002); GW distribution analysis (Dykes and Brunsdon 2007); GWPCA (Fotheringham et al. 2002; Harris et al. 2011a); GW generalised linear models (Fotheringham et al. 2002; Nakaya et al. 2005); GW discriminant analysis (Brunsdon et al. 2007; Foley and Demšar 2013); and various GW-Geostatistical hybrids (Harris et al. 2010a; 2010b; 2011b; Harris and Juggins 2011; Machuca-Mory and Deutsch 2012). Robust versions can be found in Brunsdon et al. (2002) for GW summary statistics; in Fotheringham et al. (2002), Harris et al. (2010c), Zhang and Mei (2011) for GW regression; and in Harris et al. (2011c) for GWPCA. The latter of which, we expand upon in this study.

2.3 Multivariate Outlier Detection (Global Detection)

2.3.1 Detection with Robust Mahalanobis Distances

A key concept in multivariate data analysis is to measure the similarity of objects (or observations) via some distance measure, where small distances between objects indicate that they are strongly similar (and vice-versa). Here, Mahalanobis distances (MDs) can be found that account for the size and shape of multivariate data via its covariance matrix. In the context of multivariate outlier detection, MDs can be computed from each observation vector to the data centre using

$$\mathrm{MD}_i = \left[(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right]^{0.5} \quad \text{for } i = 1, \ldots, n, \tag{1}$$

where $\mathbf{x}_i$ is the ith observation vector of dimension p; $\boldsymbol{\mu}$ is the data centre (or multivariate location), usually estimated by the arithmetic mean vector; and $\boldsymbol{\Sigma}$ is the covariance matrix. Observation vectors that are the furthest away from the data centre receive the largest MDs and are therefore most likely to be classified as outlying. Observe that a multivariate outlier is an anomalous observation vector and not one particular element in this vector. As basic estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are sensitive to outliers, for our study they need to be estimated robustly. Here, we choose the MCD estimator, whose objective is to find a subset of h observations whose basic sample covariance matrix has the lowest determinant. Crucial to the robustness and efficiency of this estimator is h, and we specify a value of $h = 0.75n$, following the recommendation of Varmuza and Filzmoser (2009, p. 43).

2.3.2 Detection with Robust PCA

For low-dimensional data, the MD approach usually suffices as a method of multivariate outlier detection, whereas for high-dimensional data (e.g. where $p > n$), problems arise in that robust estimators such as MCD cannot be used (since it needs $p < n$). One way to address this problem is to first reduce the dimensionality of the data with a PCA and then work within the PCA space to detect outliers instead. Working with reduced dimensions also saves on computational time (Filzmoser et al. 2008). PCA transforms a set of p correlated variables in to a new set of p uncorrelated variables called principal components, where dimension reduction is viable if the first few components account for most of the variation in the original data. The components are linear combinations of the original variables that follow directions of maximum variance subject to the condition of orthogonality. This transform can allow for a better understanding of differing sources of variation and key trends in the data. As outliers are a key source of variation, intuitively they should be more readily observable within the transformed PCA space than in the original data space.

For PCA, a standard result in linear algebra states that

$$\mathbf{L}\mathbf{V}\mathbf{L}^{\mathrm{T}} = \mathbf{R}, \tag{2}$$

where $\mathbf{V}$ is a diagonal matrix of eigenvalues, $\mathbf{L}$ is a matrix of eigenvectors and the matrix $\mathbf{R}$ is symmetric and positive definite. If $\mathbf{R}$ is the covariance matrix $\boldsymbol{\Sigma}$ for the $n \times p$ data matrix $\mathbf{X}$, then the eigenvalues in $\mathbf{V}$ represent the variances of the corresponding p principal components. The eigenvectors in $\mathbf{L}$ are column vectors representing the loadings of each variable on the corresponding component. It is usual to report the results for the components in decreasing order of eigenvalue (i.e. variance). If we divide each eigenvalue by $\mathrm{tr}(\mathbf{V})$, then we can report the proportion of the total variance (PTV) in the original data accounted for by each component. To use PCA as a means to detect multivariate outliers, requires $\boldsymbol{\Sigma}$ to be estimated robustly, where we again use the MCD estimator (with $h = 0.75n$).

Different types of outliers can be detected with PCA, resulting from the calculation of a score distance (SD) and an orthogonal distance (OD) at each sample location i (Hubert et al. 2005). The SD for object i is defined as

$$\mathrm{SD}_i = \sqrt{\sum_{k=1}^{q} \frac{t_{ik}^2}{v_k}}, \tag{3}$$

where $k = 1, 2, \ldots, q$ is the number of retained components; t_{ik} are the elements of the component score matrix $\mathbf{T}$, with $\mathbf{T} = \mathbf{X}\mathbf{L}_q$; and v_k is the eigenvalue of the kth component. Observe that SDs are actually MDs found within the PCA space (Varmuza and Filzmoser 2009, pp. 80–81). The OD for object i is defined as

$$\mathrm{OD}_i = \left\| \mathbf{x}_i - \boldsymbol{\mu} - \mathbf{L}_q \cdot \mathbf{t}_i^{\mathrm{T}} \right\|, \tag{4}$$

where the matrix $\mathbf{L}_q$ is a matrix of the first q eigenvectors; and $\mathbf{t}_i$ is the score vector of object i for q components. Observe that ODs reflect residuals from the PCA model and thus measure a lack of fit. Using SDs and ODs, four types of observation (vectors) can be classified, as follows:

(a) Observations that have a small SD and a small OD are not outlying and are known as regular observations.
(b) Observations that have a large SD but a small OD are outlying and are known as good leverage points. These observations are outlying when projected on the PCA space, but their residuals from the PCA model are small (i.e. they have good leverage or a strong influence on their own prediction). This outlier-type can actually stabilise a PCA fit.
(c) Observations that have a small SD but a large OD are outlying and are known as orthogonal outliers. These observations are not outlying if projected on the PCA space, but their residuals from the PCA model are large. This outlier-type can be detrimental to a basic PCA fit.
(d) Observations that have a large SD and large OD are outlying and are known as bad leverage points. This outlier-type can strongly influence a basic PCA fit, as the eigenvectors will tend to tilt toward them.

It is also possible to detect outliers from a PCA, where all p components are retained. Here, the component scores (CS) data (i.e. t_{ik}) are investigated for the first few, and the last few, components. An outlying score value for a given component at a sample location i is taken to indicate an outlying observation at that location. The rationale for this approach is that: (i) outlying observations tend to inflate variances and covariances in the first few components and (ii) for the last few components, outlying observations tend to have unusual relationships with respect to the covariance structure of data; each of which give rise to unusual CS values.

2.3.3 Determination of Cut-offs

The final step in determining whether observations are outlying or not is to specify cut-offs for the MD, SD, OD, and CS distributions. Here, an observation is deemed outlying if it has an MD, SD, OD, or CS value that is above its respective cut-off (or below, if a negative cut-off is also defined). After some experimentation, we present our study results using two different cut-off procedures for each distance measure. These can be categorised in to groups A and B, as follows:

(A) Assuming the sample data follow a multivariate normal distribution, then the squared MD data and the squared SD data, each approximately follow a chi-squared distribution with p and q degrees of freedom, respectively. Thus, for the MD and SD data, their respective cut-offs are taken as the 97.5 % quantile of the $\sqrt{\chi^2_{p,0.975}}$ and $\sqrt{\chi^2_{q,0.975}}$ distributions. For the OD data, $\mathrm{OD}^{2/3}$ is assumed approximately normal, yielding $(\mathrm{median}(\mathrm{OD}^{2/3}) + \mathrm{MAD}(\mathrm{OD}^{2/3}) \cdot z_{0.975})^{3/2}$ as a cut-off, where $z_{0.975}$ is the 97.5 % quantile of the standard normal distribution and MAD is the median absolute deviation. These cut-offs are those that are routinely defined (e.g. Varmuza and Filzmoser 2009).
(B) For the MD, SD, and OD data, their respective robust z-score data are found and the cut-offs are set at 2.5. A similar cut-off procedure is also adopted for the CS data, but where the cut-offs are set at ±2.5 to reflect outliers that correspond to large positive and large negative CS values. Here, the MD, SD, OD, and CS data are robustly standardised by subtracting their median and dividing

by their Qn scale estimator (Rousseeuw and Croux 1993). This cut-off procedure is suggested in Daszykowski et al. (2007), but for SD and OD data, only.

Observe that both cut-off procedures employ the use of robust (univariate) estimates of location and scale. Here, the mean and standard deviation have been replaced with the median and the MAD or Qn estimator, respectively. These robust estimators will down-weight the influence of outlying MD, SD, OD, and CS data.

2.4 Multivariate Spatial Outlier Detection (Local Detection)

We now describe our multivariate spatial outlier detection techniques. These techniques follow the GW methodology introduced in Sect. 2.2, resulting in robust GWMD and robust GWPCA detection methods. Unlike the calculation of MD data or a PCA, the calculation of GWMD data or a GWPCA involves regarding any observation vector $\mathbf{x}_i$ as having a certain dependence on its spatial location i with coordinates (u, v). Here, $\boldsymbol{\mu}(u, v)$ is the local mean vector and the local covariance matrix is

$$\boldsymbol{\Sigma}(u, v) = \mathbf{X}^{\mathrm{T}}\mathbf{W}(u, v)\mathbf{X}, \tag{5}$$

where $\mathbf{W}(u, v)$ is a diagonal matrix of geographic weights. For outlier detection, both $\boldsymbol{\mu}(u, v)$ and $\boldsymbol{\Sigma}(u, v)$ need to be estimated robustly; again using the MCD estimator, but now locally. We generate the weights $\mathbf{W}(u, v)$ using box-car or bi-square kernel functions, which are respectively

$$w_{ij} = 1 \quad \text{if} \quad d_{ij} \le r,\ w_{ij} = 0 \text{ otherwise}, \tag{6}$$

$$w_{ij} = \bigl(1 - (d_{ij}/r)^2\bigr)^2 \quad \text{if} \quad d_{ij} \le r,\ w_{ij} = 0 \text{ otherwise}, \tag{7}$$

where the bandwidth is the geographic distance r; and d_{ij} is the geographic distance between spatial locations of the ith and jth rows in the data matrix. For these particular kernel functions, the bandwidth is essentially the radius of a circular search window. It can be specified as: (1) a fixed distance (where the number of local observations vary within the search window) or (2) 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 fixed number of local observations is reported as a percentage of the full data set.

To find the local principal components for GWPCA, the decomposition of the local covariance matrix provides the local eigenvalues and local eigenvectors. The product of the ith row of the data matrix with the local eigenvectors for the ith location provides the ith row of local component scores. The local principal components at a location (u_i, v_i) can be written as

$$\mathbf{L}(u_i, v_i)\mathbf{V}(u_i, v_i)\mathbf{L}(u_i, v_i)^{\mathrm{T}} = \boldsymbol{\Sigma}(u_i, v_i), \tag{8}$$

where $\mathbf{L}(u_i, v_i)$ is a matrix of local eigenvectors, $\mathbf{V}(u_i, v_i)$ is a diagonal matrix of local eigenvalues, and $\boldsymbol{\Sigma}(u_i, v_i)$ is the local covariance matrix. Thus, for a GWPCA with p variables, there are p components, p eigenvalues, p sets of component loadings, and p sets of component scores at each data location.

Accordingly, the local MD, SD, OD, and CS data can be found, i.e. the data vectors $\mathrm{MD}_j(u_i, v_i)$, $\mathrm{SD}_j(u_i, v_i)$, $\mathrm{OD}_j(u_i, v_i)$, and $\mathrm{CS}_{jk}(u_i, v_i)$ at locations $i = 1, \ldots, n$;

with elements $j = 1, \ldots, N$, where N is the fixed number of observations used in each local calibration (i.e. the adaptive bandwidth size); and where the components $k = 1, 2$ and $p - 1$, p, say. This localised data is found in a fashion analogous to that defined in the global case in Sect. 2.3, where the full un-weighted data set is replaced by n GW data subsets. This implies that n sets of local MD/SD/OD/CS values are found, where each set is of size N. A multivariate spatial outlier is determined according to the size (relative to its cut-off) of those local MD/SD/OD/CS values that directly correspond to the local MD/PCA calibration (sample) point; i.e. when $j = i$ in each local MD/SD/OD/CS data set. Cut-offs for the local MD/SD/OD/CS data are the same as that described in Sect. 2.3.3, where aside from the externally-defined cut-offs of group A for MD/SD data, all other cut-offs depend on the distribution of each local MD/SD/OD/CS data set.

2.5 Key Specifications

Firstly, it is worth emphasising that the (global or local) SD and OD data and their associated cut-offs are dependent on the number of q components retained in the PCA or GWPCA model. As there is no objective choice for q, it is recommended to try with different values of q. Observe that we globally-define q, but for GWPCA, it could have been locally-defined where it would vary across space.

Secondly, for GWMD and GWPCA detection methods, the local MD, SD, OD, and CS data and their associated cut-offs depend on the bandwidth and to a lesser degree, the kernel function. Again, it is recommended to try with different bandwidths and kernels, which we demonstrate in subsequent sections. Outlier detection can be considered more locally-focused when the bi-square kernel is used, as even with a 100 % bandwidth (i.e. an adaptive bandwidth whose radii extend to all of the sample data), it still provides local detection (as weights decay with distance). If a detection method is calibrated using a box-car kernel with a 100 % bandwidth, then the corresponding global results are found (as weights are all equal to one). For box-car kernels, multivariate spatial outliers can only be detected using the smaller bandwidths.

Thirdly, the determination of cut-offs that separate background data from anomalies is not straightforward (Daszykowski et al. 2007; Filzmoser and Todorov 2013). We present our results using the cut-off procedures of Sect. 2.3.3. However, for the cut-offs of group B, we also tried basic (non-robust) z-score data. In our data simulation study of Sect. 3, we found this use of basic z-scores to sometimes improve detection performance. Thus, in a few instances, these results are reported instead of those using robust z-scores. In practise, it is recommended that both z-score options should be assessed.

3 Data Simulation Study

3.1 Simulation Algorithm

In order to objectively evaluate the GWMD and GWPCA detection methods, they are applied within a data simulation study, described by the following 21 steps and observations:

Table 1 Parameter values for the Matérn model, together with the Simple CoK means, for each of the five different variables of the simulation study

Variable number	Nugget variance	Structural variance	Correlation range (km)	Smoothing parameter	Simple CoK mean
1	0	70	27.5	2.5	25
2	0	90	27.5	2.5	50
3	0	95	27.5	2.5	45
4	0	75	27.5	2.5	30
5	0	85	27.5	2.5	40

3.1.1 Data Generation: Steps 1 to 8

1. Simulate values for five variables using an un-conditional sequential Gaussian co-simulation (e.g. Chilès and Delfiner 1999; Wackernagel 2003) using functions provided in the R *gstat* package (Pebesma 2004); where un-conditional means that the realisations are not conditioned to data. This procedure (simultaneously) generates variables that are spatially dependent and spatially co-dependent with each other. Specify a linear model of co-regionalisation (LMC) with Matérn models; themselves specified with high levels of smoothness and spatial dependence/co-dependence. The *gstat* functions ensure that all co-regionalisation matrices are positive semi-definite, which ensures that the matrix covariance function is positive definite. Values for the five variables are simulated at the 533 data locations used in the study of Sect. 4. Parameter values for the Matérn model, together with the Simple CoK means (as the co-simulation is based on this kriging form), for each of the five different variables is given in Table 1 (the cross-covariance parameter values are not given).

2. Since values are simulated for only five variables, this is low-dimensional data. Thus, following the discussions of Sect. 2.3.2, our GWPCA-based detection method that generates SD and OD data (termed GWPCA-DIST), need only show promise in detecting multivariate spatial outliers (termed local outliers) in this instance.

3. As data from any realisation are likely to be strongly spatially dependent/co-dependent from the LMC specification in step 1, it is reasonable to assume that the data are entirely free of local outliers. That is by design, all neighbouring data vectors at all 533 locations should be strongly similar in value to the data vectors at those 533 locations.

4. Multivariate (non-spatial) outliers (termed global outliers) are still possible however, so detect and mark these outliers using a non-spatial benchmark method. In this case, use a GWMD calibration with a box-car kernel, a 100 % bandwidth and a cut-off from group A; as this specification directly corresponds to the non-spatial MD approach of Filzmoser et al. (2005). An example realisation of this (un-contaminated) data with the global outliers marked is given in Fig. 2a. Observe that the outliers are highly clustered in two areas; one in an area corresponding to north–west Scotland

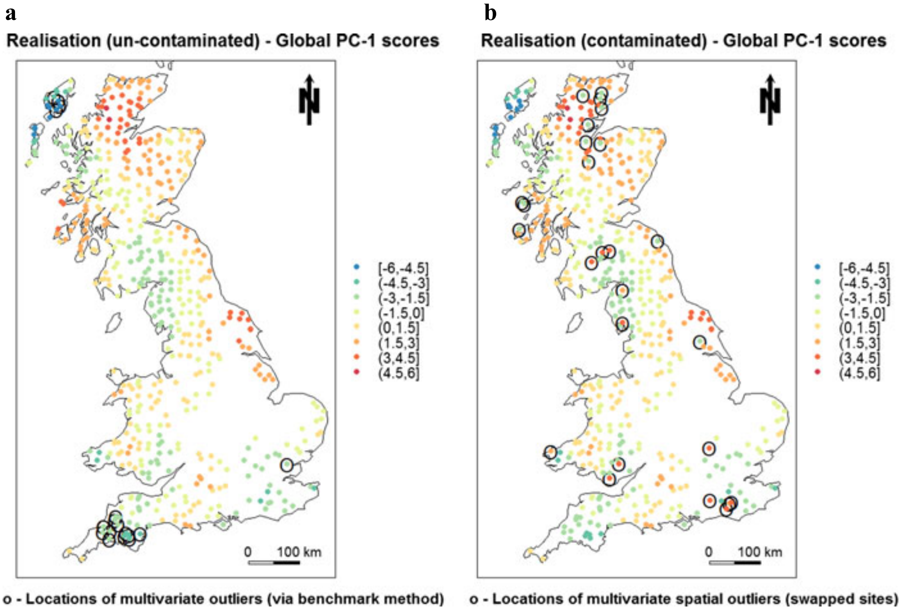
a
Realisation (un-contaminated) - Global PC-1 scores
b
Realisation (contaminated) - Global PC-1 scores
N
[-6,-4.5]
(-4.5,-3]
(-3,-1.5]
(-1.5,0]
(0,1.5]
(1.5,3]
(3,4.5]
(4.5,6]
0 100 km
o - Locations of multivariate outliers (via benchmark method)
o - Locations of multivariate spatial outliers (swapped sites)

Fig. 2 (**a**) Global PC-1 scores for un-contaminated data with (global) multivariate outliers marked; (**b**) global PC-1 scores for data that are contaminated with (local) multivariate spatial outliers (marked). Both maps are from the same realisation

and the other in an area corresponding to south–west England. This clustering of outliers is common to many realisations, highlighting a particular type of misclassification when a non-spatial detection method is naively applied (i.e. a false positive detection when an observation's spatial neighbours are similar in value, see Sect. 3.3). The number of global outliers detected for the multiple realisations of Sect. 3.2, ranged from 9 to 42 (2 % to 8 % of the data).

5. To provide a means to contaminate a realisation with local outliers, first calculate the (global) PCA scores for the first component (PC-1) of the data. Assume that the single-variable, PC-1 scores data accounts for the majority of the structure in the five-variable, realisation (see Fig. 2a).

6. Contaminate (approximately) 5 % of the realisation by swapping data at locations with high PC-1 scores with data at locations with low PC-1 scores. Thus, high and low PC-1 scores are used as an indicator for high and low levels of variation. This procedure should place individual data vectors in areas where their neighbouring data vectors are strongly dissimilar in value (i.e. locally-outlying).

7. For step 6, do not swap data at the global outlier locations identified in step 4. This ensures there are no outliers that are both global and local. This is unlikely in practise, but important for objectively evaluating the results (see steps 18 and 21).

8. For step 6, try to ensure that data clusters are not swapped with each other, as this will not produce the full complement of local outliers. For example, a worst-case scenario would involve swapping one single cluster of data with another single cluster of data elsewhere, producing data that are still locally-alike. To minimise the swapping of data clusters, the highest (97.5–100 %) and lowest (0–2.5 %) intervals of the PC-1 scores data are not used to determine the swapping locations, but instead a slightly lower (92.5–95 %) and higher (5–7.5 %) interval, respectively. An example of a contaminated realisation, where the local outliers are marked is given in Fig. 2b. Observe that there is still an element of clustering in the swapping procedure, but this is considered tolerable. Specifying lower and higher intervals for the PC-1 scores data would be counter-productive. On using this procedure, 24 or 26 local outliers are introduced.

3.1.2 Application of the GWMD/GWPCA Detection Methods: Steps 9 to 11

9. Apply the GWMD/GWPCA detection methods to a contaminated realisation assuming that the locations of all global and all local outliers are known. Observe, however, that step 8 does not ensure knowledge of the status of every data vector.

10. When applying the GWMD/GWPCA detection methods, specify them with both kernel forms and with eleven bandwidths set at 6.9 %, 10 %, 20 %, 30 %, $\ldots$, 100 % (initially, the lowest bandwidth was set at 5 %, but at a few locations a singularity occurred with the MCD estimator). This requires $2 \times 2 \times 11 = 44$ core calibrations in total, for just one realisation (2 detection methods; 2 kernels; 11 bandwidths). Further evaluations stem from the 44 calibrations, according to different cut-off procedures and the dual use of GWPCA for the GWPCA-DIST method and the GWPCA-based method that investigates the CS data (termed GWPCA-SCOR).

11. Specify GWPCA-DIST with $q = 2$ retained components. In general, experimentation with a smaller value of q improved detection performance with the OD data combined with poorer detection with the SD data. For larger values of q, the reverse was generally true. Specifying $q = 2$ is viewed a pragmatic compromise. For GWPCA-SCOR, only investigate this data from the first and last components.

3.1.3 Application of a CoK Cross-Validation Detection Method: Steps 12 to 17

12. As data are generated using un-conditional Gaussian co-simulation, it is a simple task to use its under-lying kriging model (i.e. Simple CoK) as a basis to detect outliers. In and of itself, CoK is not a method to detect outliers, but CoK cross-validation can be used. Furthermore, we calibrate CoK cross-validation with exactly the same model specifications as that used to generate the data. This enables a pseudo-robust CoK cross-validation as its matrix covariance function is not compromised by (local) outliers (i.e. it corresponds to a pre-contaminated realisation). Observe that for each realisation, spatial dependence/co-dependence is not the same as the model specified for the simulation. Only if one generated a large number of realisations and averaged the sample matrix covariance functions then that average would tend to the models specified.

13. Applying CoK cross-validation to data from a contaminated realisation, results in five leave-one-out (simultaneous) predictions, one for each variable of the data vector at a given location. Here, an outlier is identified as that which corresponds to a large residual (i.e. the actual value minus its prediction) for at least: (a) one, (b) two, (c) three, (d) four, or (e) all five variables (i.e. five variants of the detection method are investigated). For cut-offs, use the z-score procedure of group B from Sect. 2.3.3.

14. For this method, it is necessary decide on: (i) the search neighbourhood (which is commonly specified with a maximum radius, together with upper and lower bounds on the number of data locations to be used within it, e.g. Deutsch and Journel 1998) and (ii) whether raw or normalised residuals (i.e. raw residuals divided by their CoK standard errors) should be used. These decisions are inter-dependent, where the following considerations should be noted. Firstly, an outlying data vector is one that results in the matrix covariance function being a poor fit to the data. Thus, it is the parameters of the covariance functions that are key and not those of the search neighbourhood. In this respect, changing the bandwidth distance in a GW method is analogous to changing the range parameter of the covariance function (and not the radius of the search neighbourhood). Secondly, if using raw residuals, some may appear unduly large because there are few data locations within their search neighbourhoods. Normalised residuals can compensate for this, as (both kriging and) CoK standard errors reflect the number of data locations within the neighbourhood and their geometric pattern. A drawback to the use of normalised residuals is that the standard errors are poor measures of local uncertainty (e.g. Journel 1986; Goovaerts 2001), as they depend on a globally-defined matrix covariance function. Thirdly, as the CoK weights also depend on the matrix covariance function, their size will similarly reflect on the number of data locations within the neighbourhood and their geometric pattern. Except when pure nugget effect covariance functions are used, data locations that are nearest (and inside the search neighbourhood) to the prediction location receive the largest CoK weights (and thus provide the greatest influence on the prediction and in turn, the residual).

15. Considering the points of step 14, the detection performance of this method will depend on: (A) the matrix covariance function; (B) the use of raw or normalised residuals; and (C) the (three) neighbourhood specifications. As the comparison of GW methods is this article's focus, it is felt that CoK cross-validation should be specified fairly pragmatically with respect to the search neighbourhood and the residual-type. Thus a search strategy of the nearest $N = 27$ data locations within a circular neighbourhood is used (i.e. a maximum radius set to the size of sampled area with lower and upper data bounds both set to 27 or 5 % of the data), together with raw residuals. This neighbourhood is the same as that specified in the co-simulation. Future work could investigate these decisions more deeply.

16. It is, however, prudent to assess the choice of N. Here, neighbourhoods with $N > 27$, did little to alter prediction (and outlier detection) accuracy, but considerably increased computational burden. For neighbourhoods with $N < 27$, prediction and detection performance simply declined, as local information was reduced.

The use of fixed N neighbourhoods entails that CoK cross-validation with raw residuals does not suffer from varying local information within the neighbourhood; and here a preliminary investigation found this specification to consistently out-perform a specification using normalised residuals, across multiple realisations.

17. CoK cross-validation provides benchmark results from a multivariate model that is spatial by design, where intuitively, it is expected to have some success in detecting local outliers. It should provide a stringent comparative test for the GWMD/GWPCA detection methods, as its construction directly relates to the simulation study itself. Observe that the under-lying assumptions and objectives for CoK are fundamentally different from those for GWMD/GWPCA. CoK caters for stationary data relationships/structures, where spatial dependencies/co-dependencies in the data are accounted for. GWMD/GWPCA caters for non-stationary data relationships/structures, where spatial dependencies/co-dependencies are not accounted for. The objective for CoK is multivariate spatial prediction, whilst for GWMD/GWPCA, the objective is the spatial exploration of multivariate data.

3.1.4 Measuring and Displaying Detection Performance: Steps 18 to 21

18. Measure each method's detection performance by a kappa statistic that rewards for correctly identifying all local outliers and all regular observations (that are not outlying), but penalises for identifying a global outlier as outlying. The maximum value of kappa is one, which reflects a method with an ideal local outlier detection rate. The minimum kappa value is zero. In order to compare results from one realisation to another, kappa is found in a scaled form to account for a changing number of global outliers (step 4) and a changing number of local outliers (steps 5–8). For a review of related kappa statistics, see Banerjee et al. (1999).

19. For GWPCA-DIST, find kappa values that measure detection performance via: (i) a high SD value only, (ii) a high OD value only and (iii) a high SD or a high OD value. Similarly, for GWPCA-SCOR, find kappa values that measure detection performance via: (a) an outlying CS value for the first component, (b) an outlying CS value for the last component and (c) an outlying CS value for the first or last components.

20. For the GWMD/GWPCA methods, plot kappa against the range of bandwidths specified in step 10. For example, see Figs. 3, 4, 5 and 6, where boxplots of kappa values are plotted from multiple realisations; and see Figs. 7–8, where single kappa values are plotted from a single realisation. For the latter plots, kappa for the chosen CoK cross-validation variant of step 13 is presented as a dashed, dark red line.

21. Essentially, our kappa statistic is constructed so that GWMD/GWPCA methods should result in kappa values tending to one at small bandwidths and zero at large bandwidths. A GWMD/GWPCA method that performs well will detect most of the local outliers, few of the global outliers, and most of the regular observations—all at small bandwidths. At large bandwidths, a GWMD/GWPCA method should detect

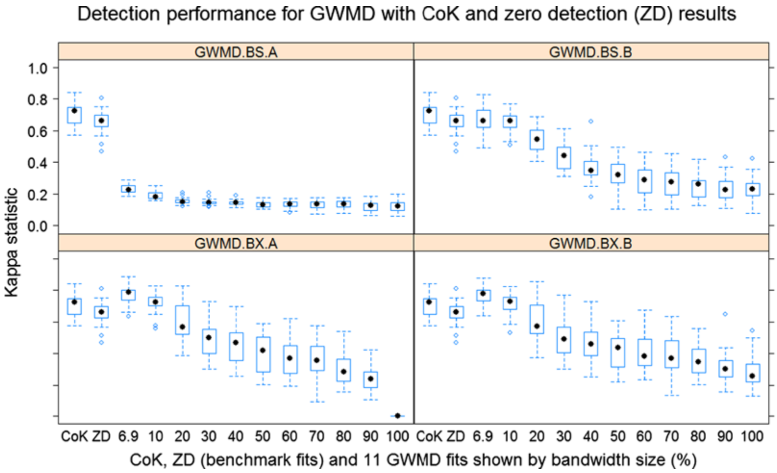

Fig. 3 Detection performance from 25 realisations: kappa boxplots versus bandwidth for GWMD, with CoK cross-validation and zero detection (ZD) results. GWMD specifications are: (**i**) bi-square kernel and group A cut-off (GWMD.BS.A); (**ii**) bi-square kernel and group B cut-off (GWMD.BS.B); (**iii**) box-car kernel and group A cut-off (GWMD.BX.A); and (**iv**) box-car kernel and group B cut-off (GWMD.BX.B). Kappa reflects local detection combined with global non-detection

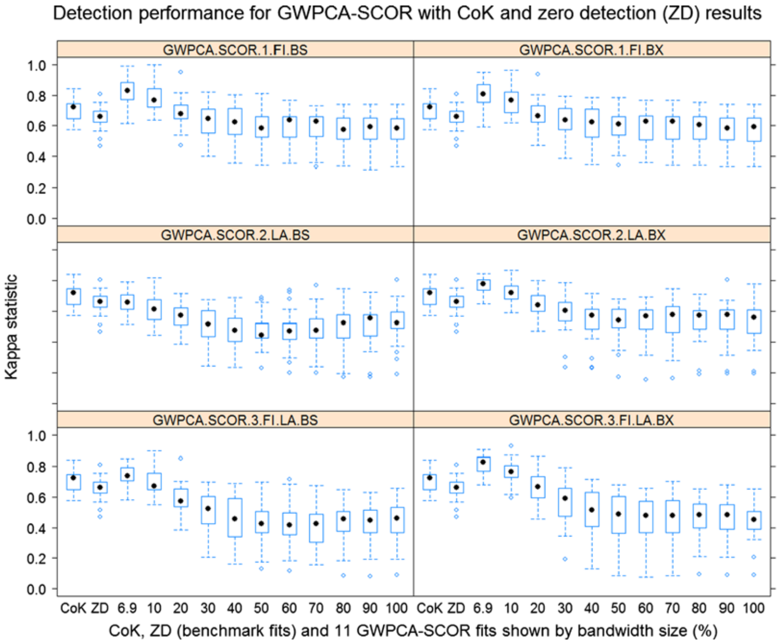

Fig. 4 Detection performance from 25 realisations: kappa boxplots versus bandwidth for GWPCA-SCOR, with CoK cross-validation and zero detection (ZD) results. GWPCA-SCOR specifications are: (**i**) bi-square kernel and first component (GWPCA.SCOR.1.FI.BS); (**ii**) box-car kernel and first component (GWPCA.SCOR.1.FI.BX); (**iii**) bi-square kernel and last component (GWPCA.SCOR.2.LA.BS); (**iv**) box-car kernel and last component (GWPCA.SCOR.2.LA.BX); (**v**) bi-square kernel and first/last component (GWPCA.SCOR.3.FI.LA.BS); and (**vi**) box-car kernel and first/last component (GWPCA.SCOR.3.FI.LA.BX)

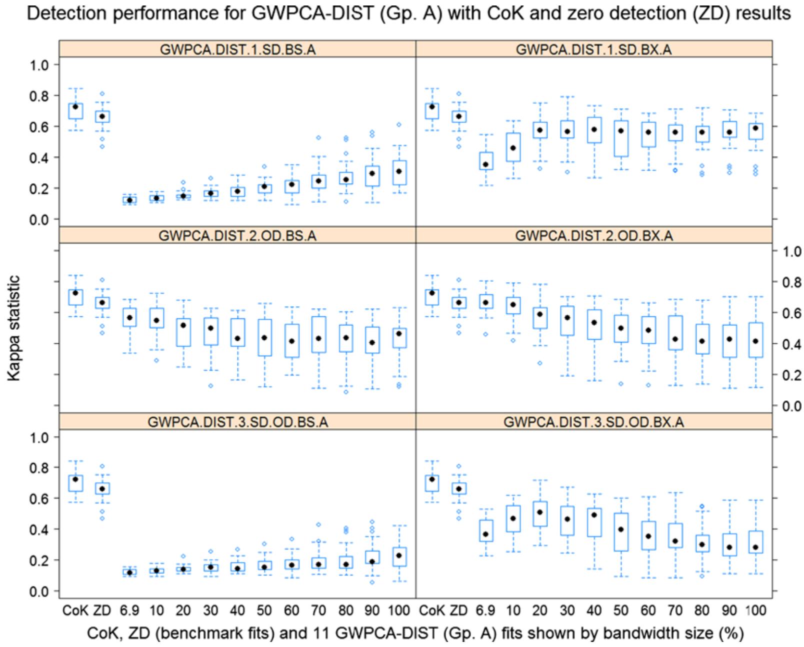

Fig. 5 Detection performance from 25 realisations: kappa boxplots versus bandwidth for GWPCA-DIST (cut-offs group A), with CoK cross-validation and zero detection (ZD) results. GWPCA-DIST specifications are: (**i**) bi-square kernel and high SD (GWPCA.DIST.1.SD.BS.A); (**ii**) box-car kernel and high SD (GWPCA.DIST.1.SD.BX.A); (**iii**) bi-square kernel and high OD (GWPCA.DIST.2.OD.BS.A); (**iv**) box-car kernel and high OD (GWPCA.DIST.2.OD.BX.A); (**v**) bi-square kernel and high SD/OD (GWPCA.DIST.3.SD.OD.BS.A); and (**vi**) box-car kernel and high SD/OD (GWPCA.DIST.3.SD.OD.BX.A)

few of the local outliers, most of the global outliers and (again) most of the regular observations. However, for each realisation, a benchmark value of kappa is needed, where a promising GWMD/GWPCA calibration is one whose kappa value is larger than this benchmark value. In this respect, a kappa value is found for a hypothetical method that results in a zero detection rate for local outliers, a 100 % non-detection (or zero detection) rate for global outliers and a 100 % detection rate for regular observations. This kappa value is referred to as zero detection, and for Figs. 7–8 is shown as dashed black line.

3.2 Detection Results from Multiple Realisations

The simulation algorithm is used to generate 25 realisations and the GWMD/GWPCA findings are summarised with kappa boxplots, yielding kappa against bandwidth trellis plots in Figs. 3–6. Kappa boxplots for zero detection (median kappa = 0.66) and the best performing CoK cross-validation variant (the one with the highest median kappa value) are also shown. For the CoK cross-validation variants, median kappa values ranged from 0.66 (for large residuals from all five variables) to 0.72 (for large residuals from at least three variables). Thus, four of the five variants provide an improvement over zero detection. Following the discussion of Sect. 2.5, basic

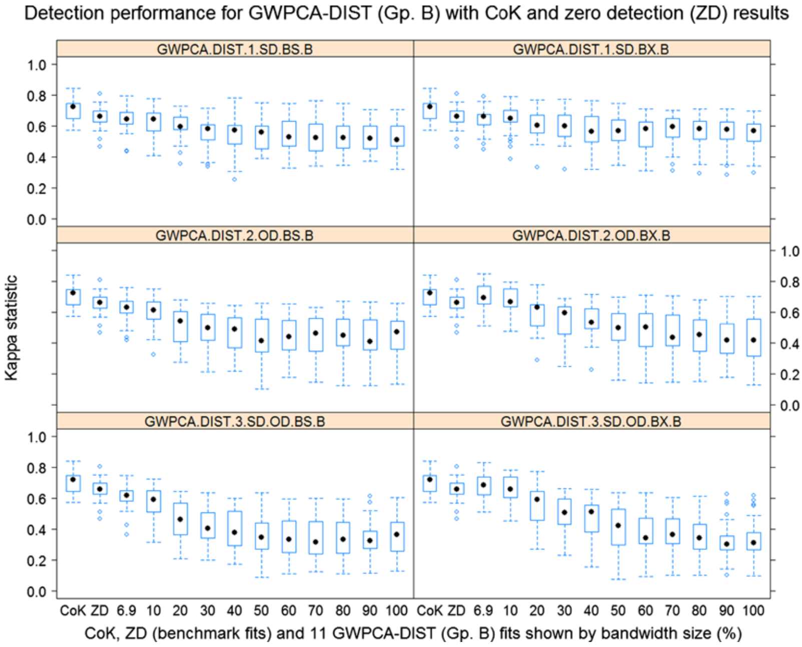

Fig. 6 Detection performance from 25 realisations: kappa boxplots versus bandwidth for GWPCA-DIST (cut-offs group B), with CoK cross-validation and zero detection (ZD) results. GWPCA-DIST specifications are: (**i**) bi-square kernel and high SD (GWPCA.DIST.1.SD.BS.B); (**ii**) box-car kernel and high SD (GWPCA.DIST.1.SD.BX.B); (**iii**) bi-square kernel and high OD (GWPCA.DIST.2.OD.BS.B); (**iv**) box-car kernel and high OD (GWPCA.DIST.2.OD.BX.B); (**v**) bi-square kernel and high SD/OD (GWPCA.DIST.3.SD.OD.BS.B); and (**vi**) box-car kernel and high SD/OD (GWPCA.DIST.3.SD.OD.BX.B)

z-scores replaced robust z-scores in the (group B) cut-off procedure for all CoK cross-validation variants and for GWMD with a bi-square kernel.

From Figs. 3–6, the following GWMD/GWPCA calibrations tend to provide better local outlier detection rates than both benchmark methods (zero detection and CoK cross-validation): (i) GWMD with a box-car kernel, using a group A or B cut-off; and (ii) all calibrations of GWPCA-SCOR, except that specified with a bi-square kernel and investigating CS data for the last component. The following GWMD/GWPCA calibrations tend to provide better local outlier detection rates than the zero detection method, but not the CoK cross-validation method: (a) GWMD with a bi-square kernel; (b) GWPCA-DIST using the SD data, specified with a box-car kernel; (c) GWPCA-DIST using the OD data, specified with a box-car kernel; and (d) GWPCA-DIST using the SD/OD data (i.e. option (iii) in step 19, above), specified with a box-car kernel (all listed methods use a group B cut-off).

The remaining GWMD/GWPCA calibrations hold no value as a method of detection, where the poorest performances are found using bi-square kernels and group A cut-offs for: (1) GWMD, (2) GWPCA-DIST using the SD data, and (3) GWPCA-DIST using the SD/OD data. Clearly, the distributional assumptions associated with group A cut-offs do not appear to hold when a bi-square kernel is specified. Furthermore, the simulation algorithm is not ideally suited to an evaluation of GWPCA-

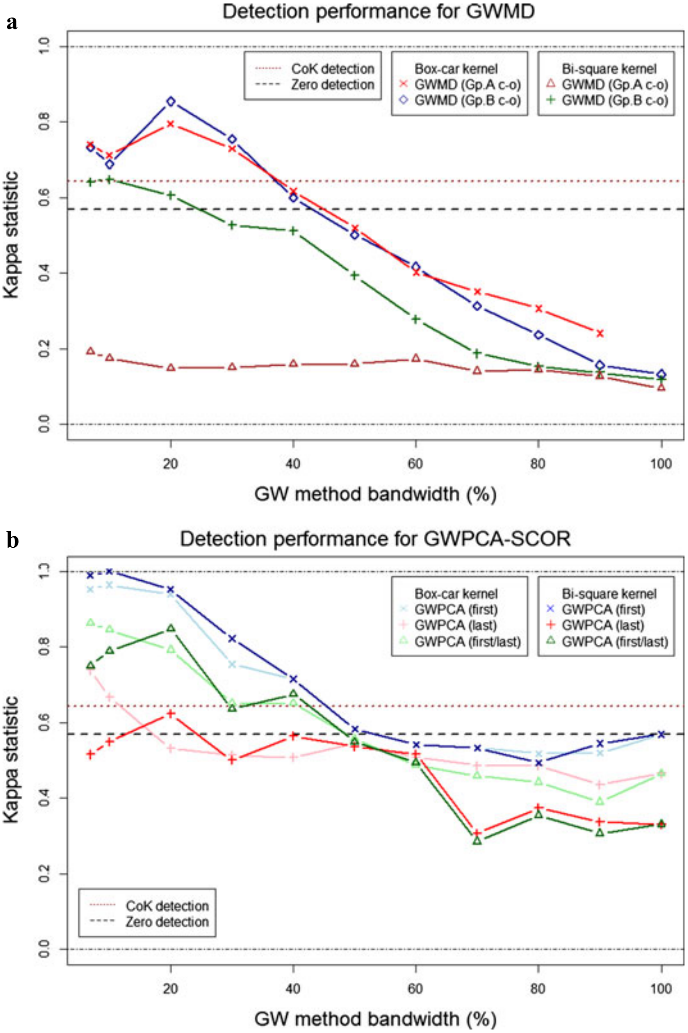

Fig. 7 Detection performance from one realisation: kappa versus bandwidth for: (**a**) GWMD and (**b**) GWPCA-SCOR

DIST. Realisations are low-dimensional with high spatial correlation. On fitting a PCA to this data, the first component commonly accounts for around 80 % of the variation and the first two components, around 90 %. The simulation of high-dimensional data sets is left for future research, as this was difficult to implement, whilst ensuring positive definiteness for the matrix covariance function. The simulation algorithms of Desbarats and Dimitrakopoulos (2000), Boucher and Dimitrakopoulos (2012) may warrant investigation, in this respect. Alternative ways to contaminate the data may also need investigation.

For the GWMD/GWPCA methods that show promise (above the zero detection line), good local outlier detection rates occur at the two smallest bandwidths of

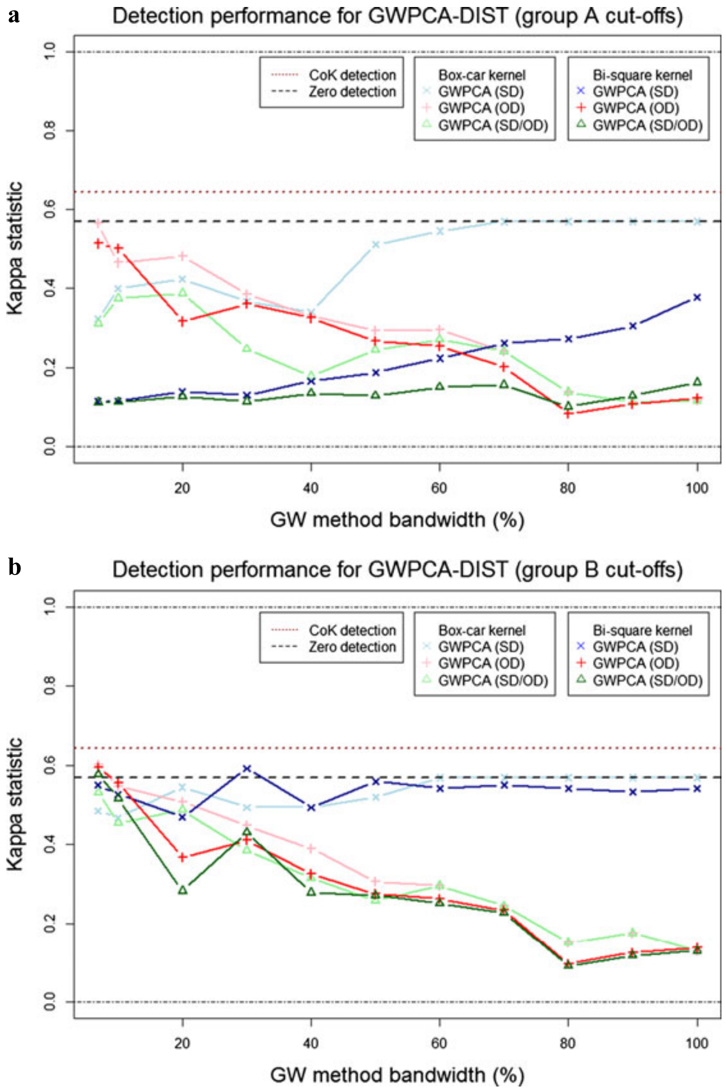

Fig. 8 Detection performance from one realisation: kappa versus bandwidth for GWPCA-DIST: (**a**) group A cut-offs and (**b**) group B cut-offs

6.9 % and 10 %. For these methods, the expected increase in kappa from large to small bandwidths is most evident in the GWMD calibrations (Fig. 3), but often only slightly so in the GWPCA-based calibrations (Figs. 4–6). Promising GWPCA-based calibrations tend to perform poorly at large bandwidths, where kappa values are relatively high, indicating poor global outlier detection rates. This in turn suggests a weak correspondence with the MD detection results of step 4, above. It is likely that a GWPCA-based method would perform more favourably in this respect, if a PCA-based detection method were used in step 4, instead. The highest median kappa value is 0.83 for GWPCA-SCOR using either the first CS data with a bi-square kernel or the first/last CS data with a box-car kernel (i.e. options (a) and (c) in step 19, above), each with a bandwidth of 6.9 % (Fig. 4). For all GWMD/GWPCA methods, the dispersion of kappa tends to reduce at the smaller bandwidths, precisely where this outcome is needed most. However, these dispersion levels tend to be larger than that found with zero detection and CoK cross-validation. The performance of CoK cross-validation is promising, although in any empirical study, such a pseudo-robust calibration would not be possible (from step 12, above). Here, a true robust calibration would need to account for an (initially) unknown set of outliers when: (A) fitting its covariance functions (see Lark 2002) and (B) predicting, say using Winsorised data (see Hawkins and Cressie 1984). A further refinement could replace the globally-defined matrix covariance function with a local version (see Haas 1996).

3.3 Detection Results from a Single Realisation

It is next useful to focus on the results from one realisation. Here, we choose a realisation (from the 25) that produced the highest kappa value. Plots of kappa against bandwidth, for this single realisation are given in Figs. 7–8. In general, the results observed for the multiple realisations also hold true for this single realisation, where GWMD and GWPCA-SCOR provide the best local outlier detection rates. Again, GWMD tends to combine good local outlier detection at small bandwidths with good global outlier detection at large bandwidths (Fig. 7a). The highest kappa value is a perfect 1 for GWPCA-SCOR, specified with a bi-square kernel and investigating CS data for the first component (Fig. 7b). This kappa occurs at a 10 % bandwidth.

Detection performance maps are given in Figs. 9b–d and 10. All maps need to be viewed in conjunction with the map in Fig. 9a, where the true locations of all local and global outliers are shown. We present the best-performing GWPCA-SCOR, GWMD, GWPCA-DIST, and CoK cross-validation calibrations, with kappa values of 1, 0.85, 0.60, and 0.64, respectively. For the GWPCA-SCOR calibration (Fig. 9b), 26 out of the 26 true local outliers are detected and there are no false positive detections. The GWMD calibration (Fig. 9c) performs well in that 24 of the 26 local outliers are detected (i.e. only two false negatives). There are, however, 12 false positives. Many false positives lie toward the edges of the realisation, where bandwidth distances will tend to be at their largest and as a result, outlier detection will not be as locally-focused as that found at other sites. The overall mis-classification rate is low, at only 14 of the 533 sites. The GWMD calibration uses a cut-off from group B, and here it is interesting to see how the same calibration performs using a cut-off from group A (Fig. 9d). Now 25 local outliers are detected, but with 20 false positives (that again tend to lie toward the edges of the realisation). The GWPCA-DIST calibration (Fig. 10a) performs poorly with only 4 local outliers detected, coupled with 6 false positive detections, resulting in an overall mis-classification rate of 28 of the 533 sites. CoK cross-validation detects 13 local outliers, but this promise is tempered by 19 false positives, one of which is a true global outlier in an area corresponding to south England (Fig. 10b). Interestingly, the false positives tend to cluster around the locations of true local outliers. This is a direct consequence of how local outliers can affect nearby CoK cross-validation predictions (which are weighted

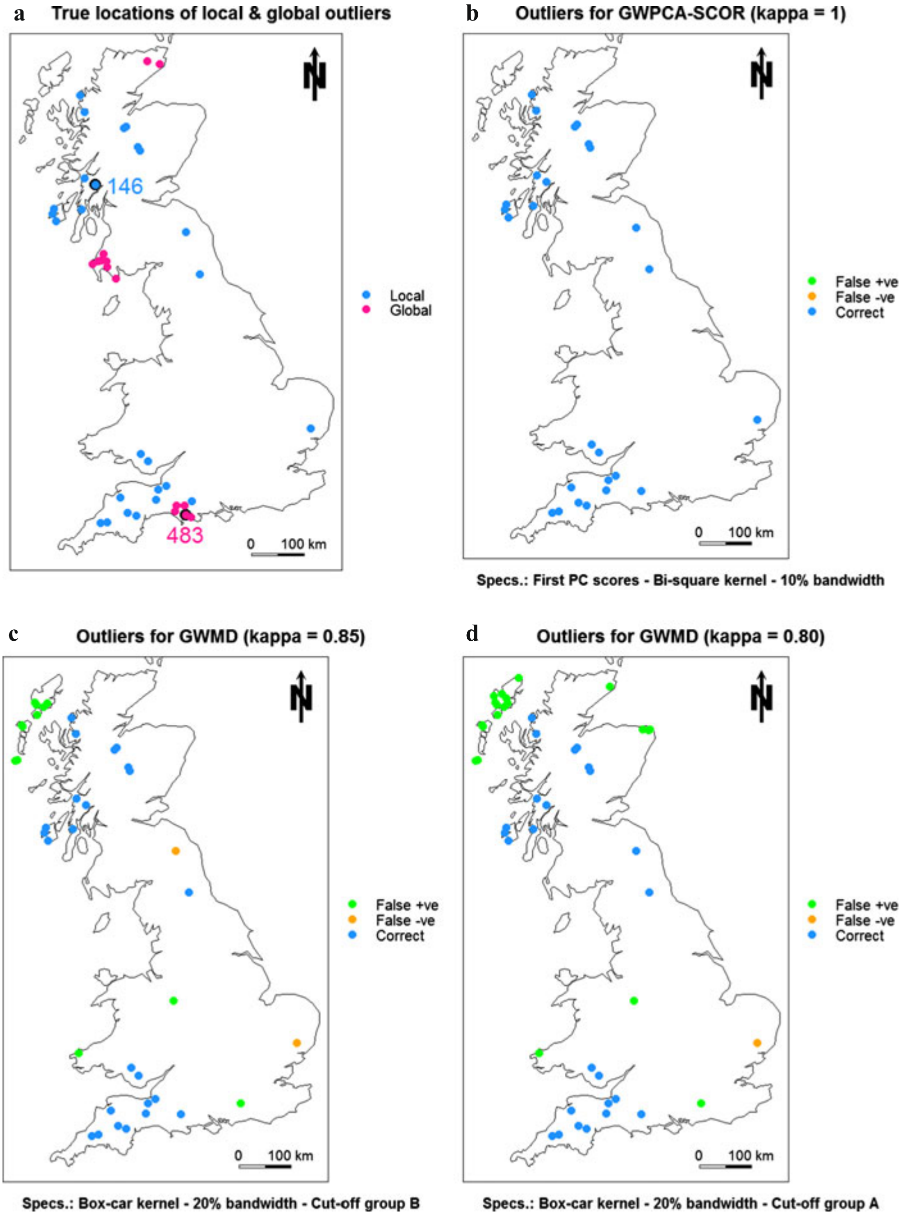
a True locations of local & global outliers
146
483
● Local
● Global
0 100 km
b Outliers for GWPCA-SCOR (kappa = 1)
● False +ve
● False -ve
● Correct
Specs.: First PC scores - Bi-square kernel - 10% bandwidth
c Outliers for GWMD (kappa = 0.85)
Specs.: Box-car kernel - 20% bandwidth - Cut-off group B
d Outliers for GWMD (kappa = 0.80)
Specs.: Box-car kernel - 20% bandwidth - Cut-off group A
N

Fig. 9 (**a**) Map of the true local and global outliers (for the two sites highlighted, see Fig. 11). Detection performance map for: (**b**) GWPCA-SCOR; (**c**) GWMD (group B cut-off); and (**d**) GWMD (group A cut-off). All results from a single realisation

means of neighbouring data), giving rise to large residuals at locations that are not themselves outlying. It appears that CoK cross-validation is able to locate the region of a local outlier, but not necessarily, its exact location. In general, the observations

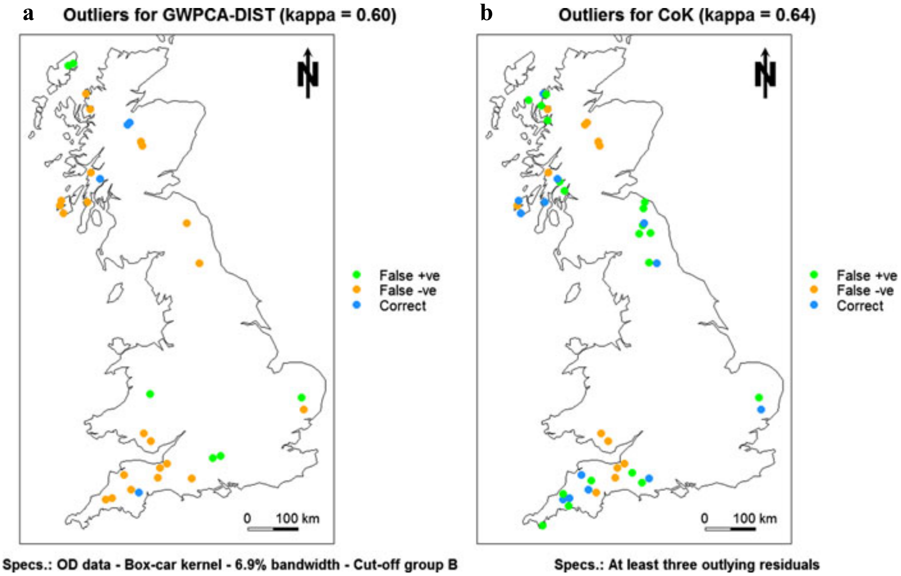
a
Outliers for GWPCA-DIST (kappa = 0.60)
N
● False +ve
● False -ve
● Correct
0 100 km
Specs.: OD data - Box-car kernel - 6.9% bandwidth - Cut-off group B
b
Outliers for CoK (kappa = 0.64)
N
● False +ve
● False -ve
● Correct
0 100 km
Specs.: At least three outlying residuals

Fig. 10 Detection performance maps for: (**a**) GWPCA-DIST; and (**b**) CoK cross-validation. All results from a single realisation

made from the performance maps for our chosen single realisation were found to be representative of all 25 realisations.

Via parallel coordinate plots (PCPlots), it is possible to visualise false positive and false negative detections with respect to the sole use of a non-spatial method, at true global and local outlier locations, respectively. Examples are given in Fig. 11. Here, at site 483 (marked in Fig. 9a), a global outlier exists. Its multivariate structure, depicted by a red line in a standard PCPlot (Fig. 11a), appears dissimilar to that found at most other sites. A different picture emerges however, when we construct a geographically weighted PCPlot (GWPCPlot) of the same data (Fig. 11b). Here, the multivariate structures at each site (except site 483) are shown as black lines with varying levels of transparency, reflecting a bi-square distance-decay weighting from site 483 (i.e. lines for the furthest away sites are essentially invisible). From the GWPCPlot, we can observe that the multivariate structure at site 483 (still depicted by a red line) is actually similar to that found at neighbouring sites. Thus, this global outlier is not a local outlier and is an example of a false positive detection in this respect. Conversely at site 146 (marked in Fig. 9a), we have a regular observation. Here, its multivariate structure can be viewed as similar to that found at a significant proportion of other sites, as depicted in the PCPlot of Fig. 11c (although it may appear outlying, it was not detected as so). However, when we construct a GWPCPlot of the same data (Fig. 11d), the multivariate structure at site 146 is highly dissimilar to that found at neighbouring sites. Thus, this regular observation is a local outlier and is an example of a false negative detection in this respect. This local outlier was detected by all five calibrations of this section (see Figs. 9–10), whereas the global outlier at site 483 was not (i.e. all detection methods performed as they should in this instance).

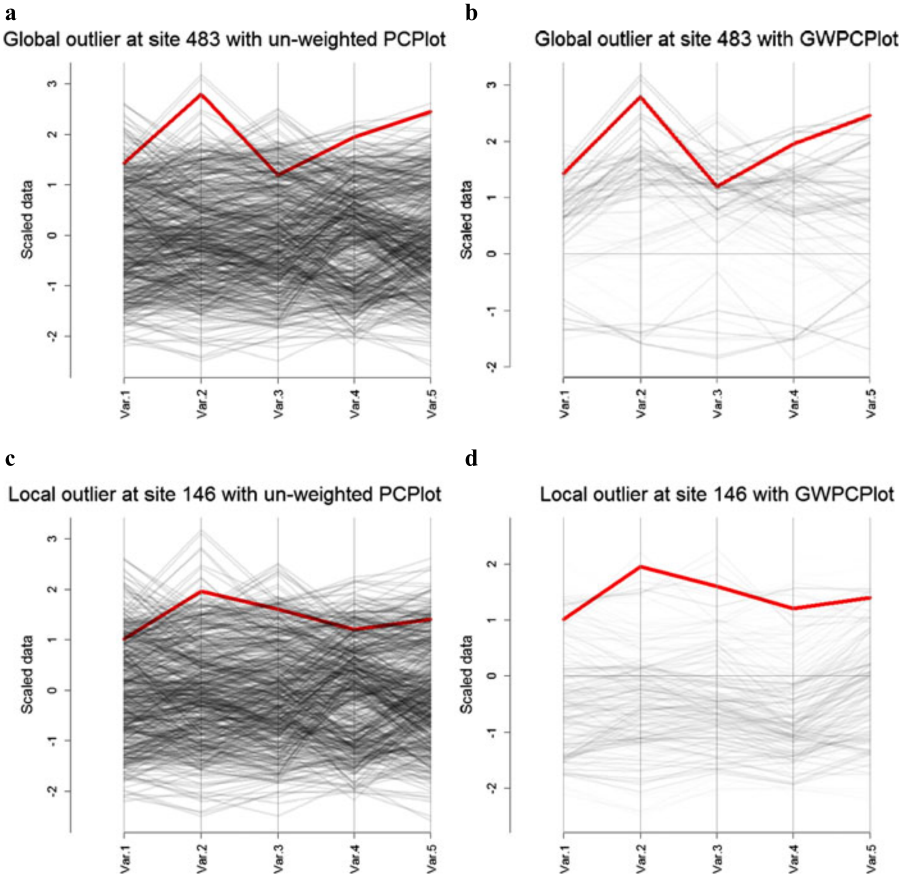

Fig. 11 PCPlots and GWPCPlots depicting examples of: a false positive (**a**)–(**b**) and a false negative (**c**)–(**d**) detections (with respect to the sole use of a non-spatial method) at global and local outlier locations, respectively. *Red line* is the multivariate structure at the example sites, 483 and 146 (see Fig. 9a). All plots from a single realisation

4 Empirical Case Study

4.1 Freshwater Chemistry Data for Great Britain

The data chosen for our empirical study is composed of eight water chemistry variables at 533 freshwater sites widely located across Great Britain. The data is a sub-set of a water chemistry sampling programme for Great Britain as part of the UK Department of Transport and Regions freshwater acidification critical loads mapping programme (Kreiser et al. 1993). Research teams within the Critical Loads Advisory Group (CLAG) then used the water chemistry data to calculate and map critical loads (CLAG CLAG Freshwaters 1995). The variables selected for this study are: pH; alkalinity (units $\mu\mathrm{eq\,L}^{-1}$, termed Alk); conductivity ($\mu\mathrm{S\,cm}^{-1}$, Cond); nitrate or NO_3^- ($\mu\mathrm{eq\,L}^{-1}$, NO3); sulphate or SO_4^{2+} ($\mu\mathrm{eq\,L}^{-1}$, SO4); phosphate or PO_4

(μeq L^{-1}, PO4); total monomeric aluminium (μg L^{-1}, AL.TM); and total organic carbon (mg L^{-1}, TOC).

4.2 Exploration with Basic and Robust GWPCA

As an example of exploring this data with a GW method, we investigate for non-stationarities in the multivariate structure of the data with basic and robust GWPCAs. As all variables aside from pH are strongly positively skewed, the GWPCAs were conducted using transformed data as well as the raw data; and in both cases, the data were then standardised. Thus seven of the eight variables were jointly transformed to approximate multivariate normality using a multivariate Box-Cox power transform (Howarth and Earle 1979; Yeo and Johnson 2000; Ruppert 2006). Cube-root, fourth-root, and log transforms were used as convenient approximations to the actual Box-Cox parameters that were found. Transformed variables are thus re-named as: Alk.T, Cond.T, NO3.T, SO4.T, PO4.T, AL.TM.T, and TOC.T.

Analysis in the transformed data space provided the clearest and most interpretable outputs; thus only these are reported. It is likely that the analysis with the raw data is compromised by the data non-normality, which is in part due to outlying observations. The use of transformed data, together with the use of an (outlier-resistant) robust GWPCA, should help mitigate against such effects. As with any GW method, the corresponding global fit is also assessed where the basic PCA results indicated PTVs for the first and the first two components combined, as 47.3 % and 66.3 %, respectively. Results for the robust PCA were similar, with PTVs for the first and the first two components combined, as 47.2 % and 67.7 %, respectively.

For basic GWPCA, an optimal bandwidth of 48.8 % is found using cross-validation, which is associated with the retention of four components. Robust GWPCA suggested a larger optimum at 92.9 %, but for comparison, we specify our robust GWPCA with the same bandwidth as that used in basic GWPCA. Bandwidth selection procedures follow that described in Harris et al. (2011a). Observe that we are now applying GWPCA in its usual guise, to explore data structure, and not to detect outliers. In this respect, the specification of an optimal (and single) bandwidth is appropriate.

For basic GWPCA, the spatial distribution of PTV for the first two components combined is given in Fig. 12a. There is clear spatial variation in the results, where both smaller and larger PTVs occur in the local case when compared to the global (PCA) case. The largest PTVs are located in south-west England and Wales, whilst the smallest PTVs are located in north-west Scotland. Observe there is a data void in central England, an area of many missing values, but also an area where freshwater acidification was not expected to be a problem. For robust GWPCA, the corresponding PTV map is given in Fig. 12b. Here, a different spatial pattern emerges to that found with basic GWPCA. Now larger PTVs always occur in the local case than in the global case. Furthermore, the largest PTVs are now also located in the eastern and northern areas of England, whilst the smallest PTVs are now more centrally located in northern Scotland. Regardless of the GWPCA specification, Scotland appears to have the most spatially-diverse water chemistry data structures. The observed differences between the basic and robust PCA/GWPCA outputs can be taken to indicate

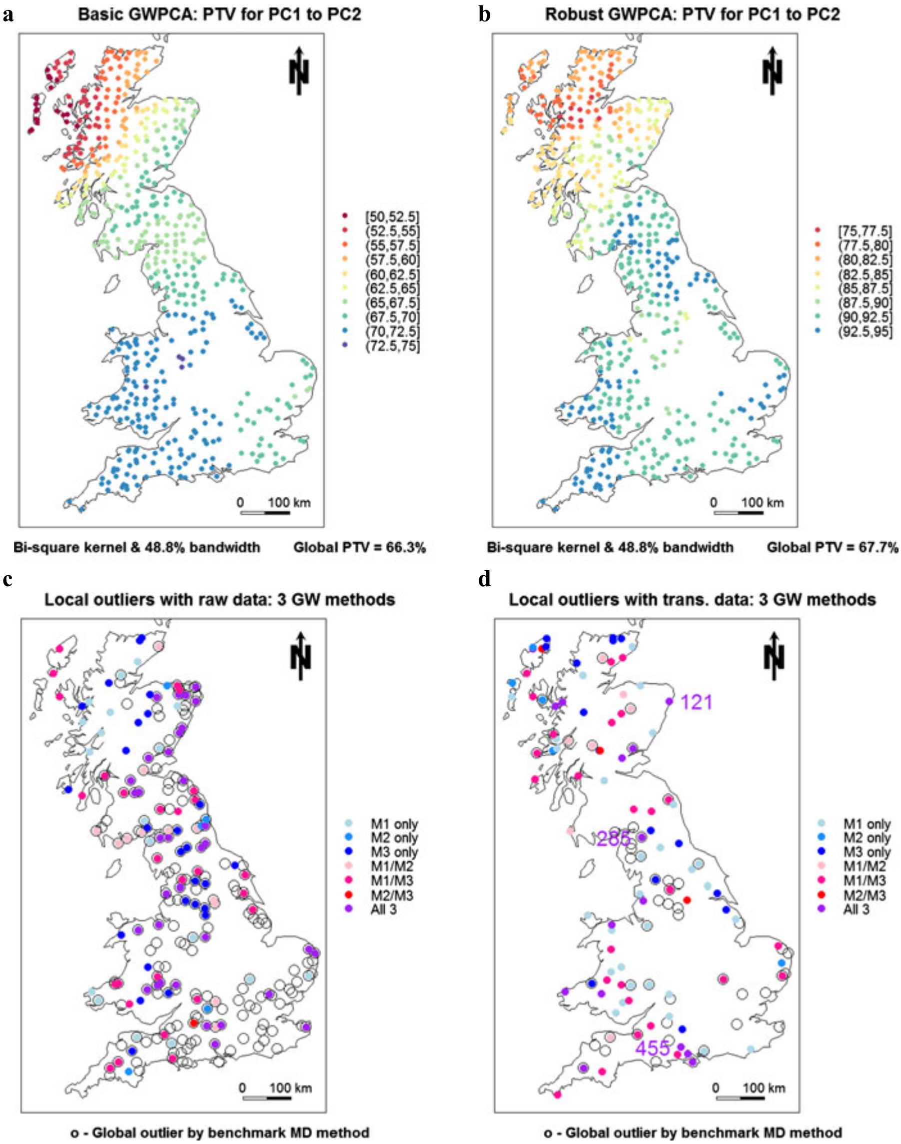

Fig. 12 (**a**) basic and (**b**) robust GWPCA PTV data for the first two components (both using transformed data); location of local outliers using three GW methods with (**c**) raw and (**d**) transformed data (where M1 is GWMD; M2 is GWPCA-DIST; M3 is GWPCA-SCOR). For (**d**), sites 121, 285, and 455 are highlighted for further scrutiny (see Fig. 13). All maps for the empirical study (see also Table 2)

the existence of many (global) multivariate outliers, some of which are likely to be locally-outlying.

Patterns in the PTV data generally relate to land cover, soil-type, or the under-lying geology; all of which are intuitively expected. For example, see the land cover, soils, and geology maps for Scotland provided by the Macaulay Land Use Research Institute (http://www.macaulay.ac.uk/ last accessed 02/9/13). Further work would need to investigate the differences between the basic and robust PTV maps more deeply, as care should be taken to ensure that robust GWPCA has not filtered out some of the key under-lying data structures. PTV data from robust GWPCA with its actual optimum bandwidth would also need exploring. Other outputs could also be investigated, such as local component loadings and local component scores.

4.3 Cautionary Notes on the Use of Data Transforms

Observe that some care must be taken when conducting a PCA with raw or transformed data, which are then used in an un-standardised or standardised form; as these data handling decisions can strongly affect analytic outputs (Baxter 1995; Cao et al. 1999). These decisions are further complicated in that the existence of outliers is often the major cause of any observed difference (Baxter 1995). Here, a data transform can both reduce and increase the number of outliers. For example, in the univariate case, with positively skewed non-zero data and a log transform, outliers in tail of the distribution are not usually outlying after the transform, whereas regular observations close to zero can be highly negative (and outlying) after the transform (Ruppert 2006). In addition, the estimated parameters of the data transform can themselves be compromised by the existence of outliers (Ruppert 2006). It follows that GWPCA will be similarly affected, but now locally, as transforms will change the spatial structure and spatial correlations in the data. In doing so, this can affect the choice of bandwidth for GWPCA, and thus alter the perception of spatial heterogeneity in the data's multivariate structure. These cautionary notes are similarly applicable to the use of GWMD and CoK with data transforms.

4.4 Global and Local Outlier Detection

Given the cautionary notes above, it is prudent to detect outliers using the raw and transformed water chemistry data. Here, we investigate the best performing GWMD, GWPCA-DIST and GWPCA-SCOR calibrations from the simulation study of Sect. 3.2, which are respectively: (i) GWMD with a box-car kernel, using a group A cut-off; (ii) GWPCA-DIST using the OD data with a box-car kernel, using a group B cut-off; and (iii) GWPCA-SCOR using the first/last CS data, also with a box-car kernel. In the simulation study, these calibrations yielded median kappa values of 0.79, 0.69, and 0.83, respectively. The determination and nature of an outlier will depend on the spatial scale at which it is viewed and in this empirical case, we choose a bandwidth of 7.5 % as our scale of investigation (i.e. the nearest 40 neighbours to each observation point). We also specify GWPCA-DIST with four retained components.

For respectively, the raw and transformed data, Figs. 12c–d map the location of potential local outliers according to our three GW methods. We also present the results from a global detection method, where we again use an MD calibration as described in step 4 of the simulation study (Sect. 3.1.1). The mapped results are also summarised in Table 2. As expected, more outliers are detected using the raw data than using the transformed data, for all four detection methods. For example, with the GWMD calibration, 91 and 77 local outliers are detected with the raw and transformed data, respectively. Of the 91 detected with the raw data, 27 remain as local

Table 2 Outlier detection results in raw and transformed data space

Method of detection	Number of raw data outliers		
		Number of transformed data outliers	
	Only outlying with raw data	Outlying with raw and transformed data	Only outlying with transformed data
GWMD	64	27	50
GWPCA-DIST	47	5	25
GWPCA-SCOR	65	23	36
All 3 GW methods in agreement	31	3	12
MD (global)	152	39	10

outliers with the transformed data. For the transformed data, 50 local outliers are detected that were not outlying with the raw data. Observe that the globally-defined transforms have the most effect on the MD calibration, where 191 and 49 global outliers are detected with the raw and transformed data, respectively. Overall, there is no strong spatial pattern or trend in the location of both global and local outliers.

If we focus on sites where all three GW methods are in agreement, then 34 and 15 local outliers are detected with the raw and transformed data, respectively (sites coloured purple in Figs. 12c–d). Of the 34 detected with the raw data, three remain as local outliers with the transformed data. For the transformed data, 12 local outliers are detected that were not outlying with the raw data. For the raw data, all 34 local outliers are also classified as global outliers. For the transformed data, three of the 15 local outliers are also classified as global outliers. For the three sites (with ID numbers of 121, 174, and 285) that are local outliers with the raw and transformed data, sites 121 and 174 are not classified as global outliers with the transformed data. Three of the 15 local outliers detected using the transformed data are highlighted in Fig. 12d. Two sites are local outliers only; one on the coast of north-east Scotland with an ID number of 121, and the other in southern England with an ID number of 455. Both sites are examples of false negative detections if only some non-spatial method of detection was applied. This can be seen in Figs. 13a–d, where both sites are not outlying with their PCPlots, whilst they are outlying with their GWPCPlots. The third highlighted site, in northern England with an ID number of 285, is a global and local outlier. The PCPlot and GWPCPlot for this site are shown in Figs. 13e–f, where the site is clearly outlying from both viewpoints.

4.5 Further Points of Interest

For our empirical study, two analytical points are worth noting: (i) the use of a data sub-set; and (ii) outlier detection with non-normal variables. With respect to the first point, the full water chemistry data set consisted of 1335 UK freshwater sites with fourteen variables. Sites for Northern Ireland and for many UK islands were removed from the analysis since the use of Euclidean distances in our GW methods may not be appropriate with these sites retained. Sites with missing values were also removed. The eight variables selected were (expertly) considered the most valuable for understanding the nature of freshwater acidification; which is our research focus. Future

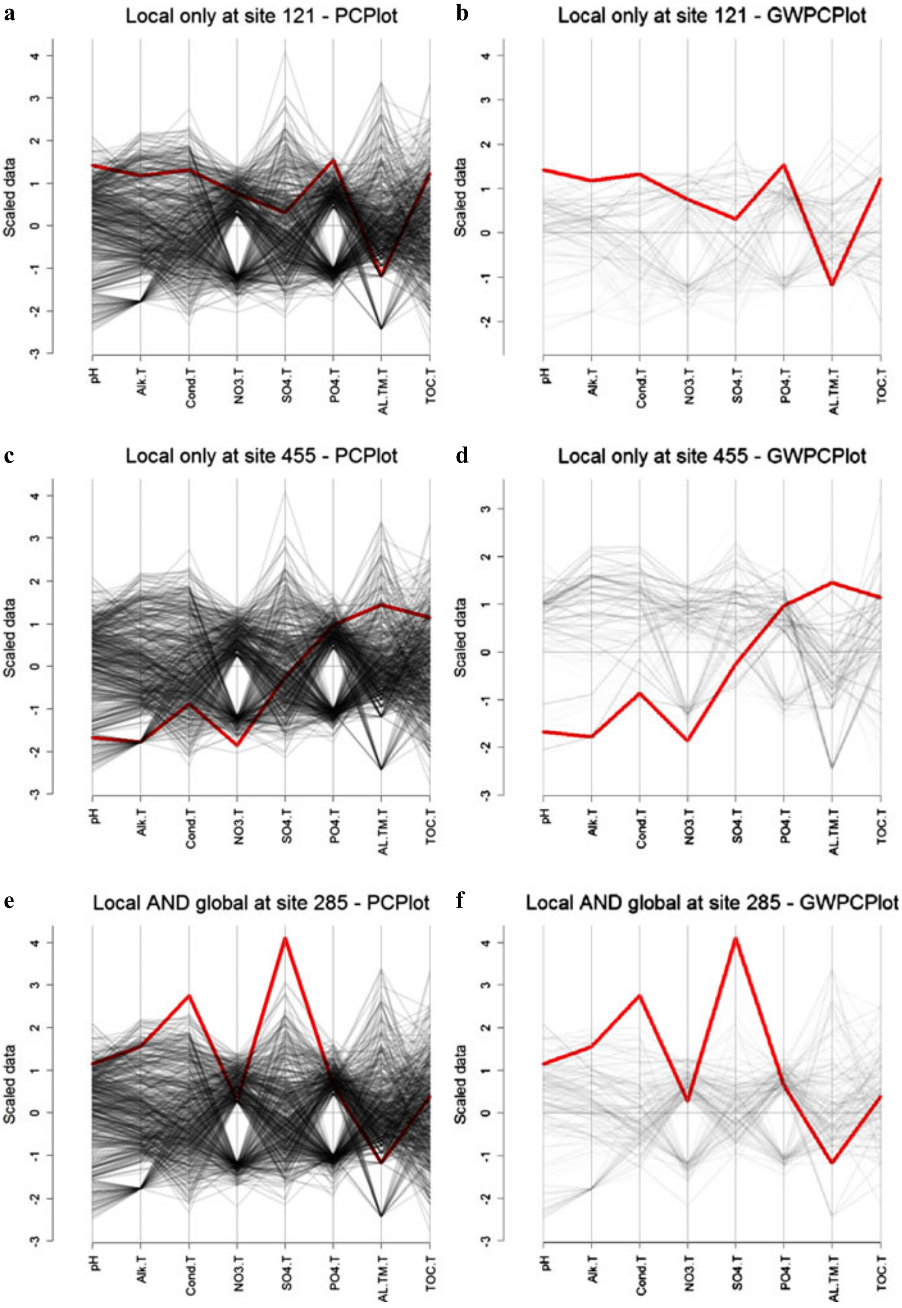

Fig. 13 PCPlots for: (**a**) a local outlier (only) at site 121; (**c**) a local outlier (only) at site 455; (**e**) a local and global outlier at site 285. GWPCPlots for: **(b)** a local outlier (only) at site 121; (**d**) a local outlier (only) at site 455; (**f**) a local and global outlier at site 285. All plots for the empirical study with transformed data (see also Fig. 12d)

work could apply the GW methods to the full UK data set, where various adaptations are likely (i.e. for use with different distance metrics and/or for use with missing data). Multivariate data with missing values is routinely encountered and will cause problems for both MDs/PCA and GWMDs/GWPCA. Here, the use of a data imputation method would be needed, such as that provided by Templ et al. (2012).

With respect to the second point, outlier detection with positively skewed data clearly poses many analytical challenges, with no simple solutions. Detection with the raw data first, then with the transformed data second, seems a pragmatic route to follow. Here, it may be worthwhile to remove or truncate the most extreme outliers that are found in the raw data investigation, prior to detections with the transformed data. In both data spaces, potential (global and/or local) outliers can be scrutinised using PCPlots and GWPCPlots. Knowledge of both raw and transformed data outliers is important, where values for the former have a direct physical interpretation. This vital property is lost in transformed data space, but if subsequent models need to be fitted using transformed data (to promote good fits) then knowledge of any potential outliers in this data space is also of value.

5 Conclusions

In this study, we have demonstrated the value of three robust geographically weighted (GW) methods for the detection of multivariate spatial outliers. One method uses local Mahalanobis distances (MDs), whilst the other two, use outputs from a local principal components analysis (PCA). All three methods perform well, both in a simulation and empirical study. Detection performance is measured both numerically and visually, using maps and (global and local) parallel coordinate plots.

Differences in detection performance primarily arise as a result of: (a) the choice of the cut-off that separates regular data from outliers and (b) the choice of kernel weighting function when calibrating a given GW method. For the methods that use local PCA, the method that investigates local component scores data (for the first few and last few components) performs better than the alternative, that investigates local MDs (within PCA space) and associated local goodness of fit distances. The latter method performs the poorest of all three GW methods.

Overall our findings are considered reasonable and worthy, where three novel outlier detection methods have been introduced and assessed. These findings are, however, dependent on the design of the simulation algorithm and the particular properties of the empirical data. Further simulation and empirical work could both endorse and enhance these findings. For example, the simulation of high-dimensional spatial data sets would complement the data simulation of this study, where only low-dimensional realisations were generated.

Acknowledgements 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. The authors gratefully acknowledge this support. We would also like to thank the anonymous reviewers whose comments helped to significantly improve this paper.

References

Anselin L (1995) Local indicators of spatial association. Geogr Anal 27:93–115

Banerjee M, Capozzoli M, McSweeney L, Sinha D (1999) Beyond kappa: a review of interrater agreement measures. Can J Stat 27:3–23

Baxter MJ (1995) Standardization and transformation in principal component analysis, with applications to archaeometry. J R Stat Soc, Ser C, Appl Stat 44:513–527

Boucher A, Dimitrakopoulos R (2012) Multivariate block-support simulation of the Yandi ore deposit. Western Australia Math Geosci 44:449–468

Brunsdon C, Fotheringham AS, Charlton M (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr Anal 28:281–298

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

Brunsdon C, Fotheringham AS, Charlton ME (2007) Geographically weighted discriminant analysis. Geogr Anal 39:376–996

Cao Y, Williams DD, Williams NE (1999) Data transformation and standardization in the multivariate analysis of river water quality. Ecol Appl 9:669–677

Chen D, Lu C, Kou Y, Chen F (2008) On detecting spatial outliers. GeoInformatica 12:455–475

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, 80 pp

Daszykowski M, Kaczmarek K, Vander Heyden Y, Walczek B (2007) Robust statistics in data analysis—a review of basic concepts. Chemom Intell Lab Syst 85:203–219

Desbarats AJ, Dimitrakopoulos R (2000) Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors. Math Geol 32:919–942

Deutsch CV, Journel AG (1998) GSLIB geostatistical software library and user's guide. Oxford University Press, New York

Dykes J, Brunsdon C (2007) Geographically weighted visualisation: interactive graphics for scale-varying exploratory analysis. IEEE Trans Vis Comput Graph 13:1161–1168

Filzmoser P, Todorov V (2013) Robust tools for the imperfect world. Inf Sci. doi:10.1016/j.ins.2012.10.017. In press

Filzmoser P, Garrett R, Reimann C (2005) Multivariate outlier detection in exploration geochemistry. Comput Geosci 31:579–587

Filzmoser P, Maronna R, Werner M (2008) Outlier identification in high dimensions. Comput Stat Data Anal 52:1694–1711

Foley P, Demšar U (2013) Using geovisual analytics to compare the performance of geographically weighted discriminant analysis versus its global counterpart, linear discriminant analysis. Int J Geogr Inf Sci 27:633–661

Fotheringham AS, Brunsdon C, Charlton ME (2002) Geographically weighted regression—the analysis of spatially varying relationships. Wiley, Chichester

Glatzer E, Müller WG (2004) Residual diagnostics for variogram fitting. Comput Geosci 30:859–866

Goovaerts P (2001) Geostatistical modelling of uncertainty in soil science. Geoderma 103:3–26

Haas TC (1996) Multivariate spatial prediction in the presence of non-linear trend and covariance non-stationarity. Environmetrics 7:145–165

Harris P, Fotheringham AS, Crespo R, Charlton M (2010a) The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. Math Geosci 42:657–680

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

Harris P, Fotheringham AS, Juggins S (2010c) Robust geographically weighed regression: a technique for quantifying spatial relationships between freshwater acidification critical loads and catchment attributes. Ann Assoc Am Geogr 100:286–306

Harris P, Juggins S (2011) Estimating freshwater critical load exceedance data for great Britain using space-varying relationship models. Math Geosci 43:265–292

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

Harris P, Brunsdon C, Fotheringham AS (2011b) Links, comparisons and extensions of the geographically weighted regression model when used as a spatial predictor. Stoch Environ Res Risk Assess 25:123–138

Harris P, Brunsdon C, Charlton M (2011c) Multivariate spatial outlier detection using geographically weighted principal components analysis. In: 7th international symposium on spatial data quality, Coimbra, Portugal

Hawkins DM, Cressie N (1984) Robust kriging—a proposal. Math Geol 16:3–18

Howarth RJ, Earle SAM (1979) Application of a generalised power transformation to geochemical data. Math Geol 11:62

Hubert M, Rousseeuw PJ, Vanden Branden K (2005) ROBPCA: a new approach to robust principal component analysis. Technometrics 47:64–79

Hubert M, Vandervieren E (2008) An adjusted boxplot for skewed distributions. Comput Stat Data Anal 52:5186–5201

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

Kou Y, Lu C-T, Chen D (2006) Spatial weighted outlier detection. In: Proceedings of the 2006 SIAM international conference on data mining, vol 614

Kreiser AM, Patrick ST, Battarbee RW (1993) Critical loads for UK freshwaters—introduction, sampling strategy and use of maps. In: Hornung M, Skeffington RA (eds) Critical loads: concepts and applications. ITE symposium no 28. HMSO, London, pp 94–98

Krige DG, Magri EJ (1982) Studies of the effects of outliers and data transformation on variogram estimates for a base metal and a gold ore body. Math Geol 14:557–564

Lark RM (2002) Robust estimation of the pseudo cross-variogram for cokriging soil properties. Eur J Soil Sci 53:253–270

Liu H, Jezek KC, O'Kelly M (2001) Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS. Int J Geogr Inf Sci 15:721–741

Ljung GM (1993) On outlier detection in time series. J R Stat Soc B 55:559–567

Lu C-T, Chen D, Kou Y (2004) Multivariate spatial outlier detection. Int J Artif Intell Tools 13:801–811

Machuca-Mory DF, Deutsch CV (2012) Non-stationary geostatistical modeling based on distance weighted statistics and distributions. Math Geosci 45:31–48

Nakaya T, Fotheringham AS, Brunsdon C, Charlton M (2005) Geographically weighted Poisson regression for disease association mapping. Stat Med 24:2695–2717

Pebesma EJ (2004) Multivariate geostatistics in S: the gstat package. Comput Geosci 30:683–691

Rousseeuw PJ (1985) Multivariate estimation with high breakdown point. In: Grossman W, Pflug G, Vincze I, Wertz W (eds) Mathematical statistics and applications, vol B. Reidel, Dordrecht, pp 283–297

Rousseeuw PJ, Croux C (1993) Alternatives to median absolute deviation. J Am Stat Assoc 88:1273–1283

Rousseeuw PJ, Ruts I, Tukey JW (1999) The bagplot: a bivariate boxplot. Am Stat 53:382–387

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

Ruppert D (2006) Multivariate transformations. In: Encyclopedia of environmetrics. Wiley, New York

Sun Y, Genton M (2011) Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. Environmetrics 23:54–64

Templ M, Alfons A, Filzmoser P (2012) Exploring incomplete data using visualization tools. J Adv Data Anal Class 6:29–47

Varmuza K, Filzmoser P (2009) Introduction to multivariate statistical analysis in chemometrics. CRC Press, Boca Raton

Wackernagel H (2003) Multivariate geostatistics—an introduction with applications. Springer, Berlin

Wheeler D (2007) Diagnostic tools and a remedial method for collinearity in geographically weighted regression. Environ Plan A 39:2461–2481

Yeo I, Johnson R (2000) A new family of power transformations to improve normality or symmetry. Biometrika 87:954–959

Zhang H, Mei C (2011) Local least absolute deviation estimation of spatially varying coefficient models: robust geographically weighted regression approaches. Int J Geogr Inf Sci 25:1467–1489