geographical analysis

Geographical Analysis (2015) **47**, 146–172

Enhancements to a Geographically Weighted Principal Component Analysis in the Context of an Application to an Environmental Data Set

Paul Harris[1], Annemarie Clarke[2], Steve Juggins[3], Chris Brunsdon[4], Martin Charlton[4]

[1]Sustainable Soils and Grassland Systems, Rothamsted Research, Okehampton, Devon, U.K., [2]APEM Ltd, Llantrisant, U.K., [3]School of Geography, Politics and Sociology, University of Newcastle, Newcastle upon Tyne, U.K., [4]National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth, Ireland

In many physical geography settings, principal component analysis (PCA) is applied without consideration for important spatial effects, and in doing so, tends to provide an incomplete understanding of a given process. In such circumstances, a spatial adaptation of PCA can be adopted, and to this end, this study focuses on the use of geographically weighted principal component analysis (GWPCA). GWPCA is a localized version of PCA that is an appropriate exploratory tool when a need exists to investigate for a certain spatial heterogeneity in the structure of a multivariate data set. This study provides enhancements to GWPCA with respect to: (i) finding the scale at which each localized PCA should operate; and (ii) visualizing the copious amounts of output that result from its application. An extension of GWPCA is also proposed, where it is used to detect multivariate spatial outliers. These advancements in GWPCA are demonstrated using an environmental freshwater chemistry data set, where a commentary on the use of preprocessed (transformed and standardized) data is also presented. The study is structured as follows: (1) the GWPCA methodology; (2) a description of the case study data; (3) the GWPCA application, demonstrating the value of the proposed advancements; and (4) conclusions. Most GWPCA functions have been incorporated within the GWmodel R package.

Introduction

Principal component analysis (PCA) is a core method for multivariate analysis in many areas of human (e.g., Johnston 1978; Griffith and Amrhein 1997) and physical (e.g., Mather 1976; Legendre and Legendre 1998; Davis 2002) geography, where for this study we focus on its use in the latter setting. A member of the unconstrained ordination family, PCA transforms a set of *m*

Correspondence: Paul Harris, Sustainable Soils and Grassland Systems, Rothamsted Research, North Wyke, Okehampton, Devon EX20 2SB, U.K.
e-mail: paul.harris@rothamsted.ac.uk

Submitted: February 18, 2013. Revised version accepted: January 28, 2014.

doi: 10.1111/gean.12048

correlated variables into a new set of *m* uncorrelated variables called components. The components are linear combinations of the original variables and can allow for a better understanding of differing sources of variation and key trends in data. These trends can be visualized and interpreted using associated graphics. Its use as a dimension reduction technique is viable if the first few components account for most of the variation in the original data. In an ecological setting, common applications of PCA are to environmental data sets (e.g., the soils biogeochemistry data in Kaspari and Yanoviak 2009), although via a suitable transform, PCA can also be applied to species abundance data (Legendre and Gallagher 2001).

In geographical settings, PCA often is applied without consideration for important spatial effects (e.g., see Demšar et al. 2013). Such naive applications can be problematic because spatial effects often provide a more complete understanding of a given process. In this respect, a common approach is to adapt PCA to account for some form of spatial autocorrelation effect (Wartenberg 1985; Dray, Legendre, and Peres-Neto 2006; Griffith and Peres-Neto 2006; Jombart et al. 2008; Jombart, Dray, and Dufour 2009). Here, a spatial weighting matrix that reflects the interaction between spatial units is needed so that spatial autocorrelation can be measured, say using Moran's *I*. Alternatively, and the subject of this study, (the global) PCA can be replaced with a (local) geographically weighted principal components analysis (GWPCA) (Fotheringham, Brunsdon, and Charlton 2002, pp. 196–202), when we want to account for certain spatial heterogeneity in data. Whereas PCA with autocorrelation effects models second-order, variance spatial effects, GWPCA models first-order, mean response spatial effects. Commonly, the former is calibrated in a stationary form, while the latter is a nonstationary model. Both approaches have merit, and if compared, may fit the same data set equally as well. However, the interpretation of the fit is the important feature in context of: (i) the spatial characteristics of the data being modeled; and (ii) the research question being asked.

Such conceptual differences are analogous to the use of regression with a spatially autocorrelated error term (e.g., Schabenberger and Gotway 2005) or GW regression (GWR) (Brunsdon, Fotheringham, and Charlton 1996, 1998) when choosing a regression model for spatially referenced data. Example applications or discussions of GWR models in physical geography (or spatial ecology) can be found in Brunsdon, MaClatchey, and Unwin (2001), Atkinson et al. (2003), Jetz, Rahbak, and Lichstein (2005), Austin (2007), Miller, Franklin, and Aspinall (2007), Foody (2008), and Harris, Fotheringham, and Juggins (2010), where GWR is recommended for exploring or investigating spatial heterogeneity in data relationships, but not necessarily for inference (Jetz, Rahbak, and Lichstein 2005; see also Wheeler and Tiefelsdorf 2005; Páez, Farber, and Wheeler 2011). GWPCA extends the multivariate GW modeling paradigm in that, unlike GWR, no model is specified (i.e., there is no predefined response and predictor variable division).

In GWPCA, a different localized PCA is computed at target locations, and as such, the results vary continuously over space, allowing them to be mapped. This visualization permits a local identification of any change in structure of a multivariate data set, pinpointing locations where results from the global PCA are inappropriate or oversimplistic. This identification, in turn, allows for better informed model decisions for any analysis that may follow, such as a clustering or regression analysis when orthogonal input data are required. Key challenges in GWPCA include: (a) finding the scale at which each localized PCA should operate; and (b) visualizing and interpreting the copious output that results from its application. Both of these issues have been addressed, at least in part, in Harris, Brunsdon, and Charlton (2011).

For this study, we provide useful and novel enhancements to both challenges through the respective use of: (1) a robust bandwidth selection procedure; and (2) visualizations that stem from the application of a clustering algorithm to GWPCA output. An extension of GWPCA is also introduced; one used to detect multivariate spatial outliers. In addition to the three methodological advancements, a fourth study objective is the illustration of a GWPCA from a physical geography perspective. Here, we use a freshwater chemistry data set for lakes and headwaters collected over all of Great Britain (CLAG Freshwaters 1995).

Given that variation in the study data is largely driven by the (anthropogenic and natural) deposition of various (acidifying and nonacidifying) compounds, types of land use, and geology, all of which vary considerably across Great Britain, the structures and relationships in this data set are expected to vary similarly across space. Here, a GWPCA allows us to uncover such changes in a single, coherent analysis, which in turn can provide valuable regional insights into important environmental concerns, such as acidification (Hornung et al. 1995) and eutrophication (Pretty et al. 2003), and their potentially harmful impacts on freshwater life. For this study, our analyses focus on acidification.

The nature of the study data is such that they are first transformed, and then they are standardized (i.e., centered and scaled), prior to the GWPCA. As with PCA, analytical output from GWPCA is strongly dependent on such data preprocessing decisions, whereas for GWPCA, whether these adjustments should be conducted globally or locally is unclear. In this respect, a useful commentary on this important topic is also presented. Thus, for methodological advances and application, our study is intended to complement that of Harris, Brunsdon, and Charlton (2011), which illustrated advances in GWPCA from a human geography perspective. Outside of the methodological presentations of Fotheringham, Brunsdon, and Charlton (2002) and Harris, Brunsdon, and Charlton (2011), the only known applications of GWPCA can be found in Lloyd (2010) and in Kumar, Lal, and Lloyd (2012) using population census and soils geochemistry data, respectively.

GWPCA methodology

GWPCA uses a moving window weighting approach, for which localized components are found at target locations. For an individual PCA at a target location, we weight all neighboring observations according to the characteristics of some kernel weighting function, and then locally apply standard PCA to these weighted data. The size of the window over which this localized PCA might apply is controlled by the kernel's bandwidth. Small bandwidths lead to more rapid spatial variation in the results whereas large bandwidths yield results increasingly close to the global PCA solution.

GWPCA

More formally, if spatial location i has coordinates (u, v), then GWPCA involves a vector of observed variables $\mathbf{x}_i$ being conceptualized as having a certain dependence on its location i, where $\mu(u_i, v_i)$ and $\Sigma(u_i, v_i)$ are the local mean vector and the local variance-covariance matrix, respectively. The local variance-covariance matrix is

$$\boldsymbol{\Sigma}(u_i, v_i) = \mathbf{X}^{\mathrm{T}}\mathbf{W}(u_i, v_i)\mathbf{X}, \tag{1}$$

where $\mathbf{X}$ is the data matrix (with n rows for the observations, and m columns for the variables), and $\mathbf{W}(u_i, v_i)$ is a diagonal matrix of geographic weights, which for this study are generated using one of two particular kernel functions. To find the local principal components at location (u_i, v_i), the decomposition of the local variance-covariance matrix provides the local eigenvalues and local eigenvectors (or loading vectors) with

$$\mathbf{L}(u_i, v_i)\mathbf{V}(u_i, v_i)\mathbf{L}(u_i, v_i)^{\mathrm{T}} = \mathbf{\Sigma}(u_i, v_i), \tag{2}$$

where $\mathbf{L}(u_i, v_i)$ is a matrix of local eigenvectors, and $\mathbf{V}(u_i, v_i)$ is a diagonal matrix of local eigenvalues. A matrix of local component scores $\mathbf{T}(u_i, v_i)$ can be found using

$$\mathbf{T}(u_i, v_i) = \mathbf{XL}(u_i, v_i), \tag{3}$$

where the product of the ***i***th row of the data matrix with the local eigenvectors for the ***i***th location provides the ***i***th row of local component scores. If we divide each local eigenvalue by tr($\mathbf{V}(u_i, v_i)$), then we find localized versions of the proportion of the total variance (PTV) in the original data accounted for by each component. Thus, at each observed location for a GWPCA with m variables, there are m components, m eigenvalues, m sets of component loadings (each of size $m \times m$), and m sets of component scores (each of size $n \times m$). We can also obtain eigenvalues and their associated eigenvectors at unobserved locations, although as no data exist for these locations, we cannot obtain component scores. In numerical ecology, component loadings and component scores are commonly referred to as variable (or species) scores and site scores, respectively. Component loadings are the correlation coefficients between the component scores and the raw data.

Kernel weighting functions

At any GW method calibration stage, experimenting with diverse kernel weighting functions is prudent. In this study, box-car and bi-square kernels are used, where the former relates to an unweighted moving window while the latter relates to a weighted one. These kernels, respectively, are defined as

$$w_{ij} = 1 \text{ if } d_{ij} \le r \text{ and } w_{ij} = 0 \text{ otherwise}, \tag{4}$$

$$w_{ij} = \left(1 - \left(d_{ij}/r\right)^2\right)^2 \text{ if } d_{ij} \le r \text{ and } w_{ij} = 0 \text{ otherwise}, \tag{5}$$

where the bandwidth is the geographic distance r; d_{ij} is the geographic distance between spatial locations of the ***i***th and ***j***th rows in the data matrix; and w_{ij} is the geographic weight attached to an observation point indexed by j, for a calibration (in this study, an observation) point indexed by i. Bandwidths can be specified either as a fixed distance or as a fixed number of local observations (i.e., an adaptive distance). For this study, only the latter are specified because they suit the irregular sample configuration of the study data. These are reported as a percentage of the full data set.

Commonly, a GW method is specified with a distance decay kernel, such as the bi-square, because this specification tends to provide outputs that vary smoothly over space. Furthermore, such a specification still can provide a local form while using (and benefitting from) all of the sample data at each target location. Conversely, the use of a box-car kernel is more likely to provide outputs that appear discontinuous over space, which can be useful where the detection of outlying relationships is more likely (Lloyd and Shuttleworth 2005; Harris and Brunsdon 2010).

Box-car calibrations also are useful, in that if a 100% bandwidth is specified, the corresponding global model is found. The spatial process may be essentially homogeneous, in which case, an application of the global model suffices. Here, the use of a GW method can provide a worthy confirmation of this homogeneity.

Basic and robust bandwidth selection

A key challenge in GWPCA is finding the scale at which each localized PCA should operate; that is, choosing the kernel bandwidth. In most GWPCA studies to date, this bandwidth has been user specified (Fotheringham, Brunsdon, and Charlton 2002; Lloyd 2010; Kumar, Lal, and Lloyd 2012). For this study, we are guided by an automatic routine for bandwidth selection, as described in Harris, Brunsdon, and Charlton (2011), which is similar in spirit to that used for bandwidth selection in GWR via cross-validation (e.g., Brunsdon, Fotheringham, and Charlton 1998). We also provide an enhancement to this procedure that allows basic and robust forms. Our robust procedure has similar objectives to robust bandwidth selection in GWR (Farber and Páez 2007).

To describe how the bandwidth is selected in GWPCA, we first discuss properties of PCA. If there are m variables in the data matrix $\mathbf{X}$, so that each observation is a vector in m-dimensional space, the component scores corresponding to components $q + 1$ to m represent the Euclidean distances along the axes of the corresponding orthogonal vectors to a q-dimensional linear subspace. Here, the q-dimensional subspace is spanned by the first q loadings (also m-dimensional vectors), and is the subspace that maximizes the variance of the data points projected onto that subspace. For dimension reduction, q is chosen so that this subspace contains a high PTV, and thus components $q + 1$ to m represent the deviation from this subspace. In terms of component scores, the first q components are described by $\mathbf{XL}_q$, and the remaining components by $\mathbf{XL}_{(-q)}$ (where $\mathbf{L}_q$ denotes the loading matrix $\mathbf{L}$ with all but the first q columns removed, and $\mathbf{L}_{(-q)}$ denotes $\mathbf{L}$ with the first q columns removed). Jolliffe (2002) shows that the best (least squares) rank q approximation to $\mathbf{X}$ is $\mathbf{XL}_q\mathbf{L}_q^{\mathrm{T}}$, and that the residual matrix $\mathbf{S}$, given by $\mathbf{S} = \mathbf{X} - \mathbf{XL}_q\mathbf{L}_q^{\mathrm{T}}$, can also be written as $\mathbf{S} = \mathbf{XL}_{(-q)}\mathbf{L}_{(-q)}^{\mathrm{T}}$. Thus, via principal components, we aim to find the minimum of the expression

$$\sum_{ik}([\mathbf{X}]_{ik} - [\mathbf{S}]_{ik})^2 \tag{6}$$

with respect to $\mathbf{S}$, where $\mathbf{S}$ is a rank q matrix; and the problem is solved with expression (6). Therefore, the variance levels of the components of $\mathbf{S}$ measure the "goodness of fit" (GOF) of the projected subplanes, where

$$\mathrm{GOF}_i = \sum_{k=q+1}^{k=m} s_{ik}^2 \tag{7}$$

is the GOF for the ***i***th observation, and s_{ik} is the ***k***th component score for observation i; that is, the ***ik***th element of $\mathbf{S}$. The total GOF for the entire data set is

$$\mathrm{GOF} = \sum_{i=1}^{i=n} \mathrm{GOF}_i. \tag{8}$$

For GWPCA, the local principal components for the ***i***th location represent a similar projection, but with the corresponding loadings defined locally. That is, we now find $\mathbf{S}$ to minimize

$$\sum_{ik} w_i\left([\mathbf{X}]_{ik} - [\mathbf{S}]_{ik}\right)^2, \tag{9}$$

where w_i is a locally defined weight for location i. GOF statistics for GWPCA can be defined in an analogous fashion to those for PCA, except that in each locality, **S** is defined using local weights. In turn, a total GOF statistic provides a means of finding an optimal bandwidth for GWPCA, where we use a "leave-one-out" method when computing the terms of the statistic. Here, a "leave-one-out" total GOF statistic can be computed for all possible bandwidths, and an optimal bandwidth relates to the smallest total GOF value found. The number of components to retain (i.e., the value of q) is decided upon a priori. An optimal bandwidth cannot be found if we wish to retain all m components; in this case, the bandwidth must be user specified. We recommend that bandwidth selection results be investigated for a range of values of q before deciding on one particular value.

Finally, an equivalent measure to this total GOF statistic (or cross-validation score) is possible if we take the mean of the "leave-one-out" GOF data (termed leave-one-out residuals, LOORs), which when computed for all possible bandwidths results in the same optimum. In turn, if we take the median of the LOOR data and compute for all possible bandwidths, an alternative optimum can be found, which should be robust to the influence of (outlying) observations that contribute the most to the total GOF statistic.

Using GWPCA to detect multivariate spatial outliers

Outlier detection is often a key task in a statistical analysis. It can be used not only as a data cleaning or screening exercise, but also to uncover interesting or unusual properties in the data that have not been considered before. For multivariate data sets, Mahalanobis distances (MDs) can be used to detect multivariate outliers (Filzmoser, Garrett, and Reimann 2005), where MDs account for the size and shape of multivariate data via their covariance matrix. Output from a PCA can also be used to detect multivariate outliers (Hubert, Rousseeuw, and Vanden Branden 2005; Filzmoser, Maronna, and Werner 2008), where in the transformed PCA space, outliers may be more readily observable. However, MD and PCA approaches are themselves sensitive to outliers, and as such, can compromise a basic calibration prior to its use as a method of detection. In this respect, robust MD and PCA detections are recommended (e.g., Rousseeuw et al. 2006; Daszykowski et al. 2007; Varmuza and Filzmoser 2009, pp. 47–50, 66–69, 78–81).

For spatial applications, MD- and PCA-based methods can only detect multivariate outliers in a nonspatial manner. This can result in a false-positive detection, such as when an observation's spatial neighbors are similar in value (i.e., the observation is not spatially outlying), or a false-negative detection when its spatial neighbors are dissimilar in value (i.e., the observation is spatially outlying). It follows that a GW MD- and a GWPCA-based methods could be constructed to detect outliers, so that these misclassifications are addressed. Both GW methods would detect multivariate spatial outliers, where their basic forms would need to be replaced with robust versions. Initial work on the detection of such outliers can be found in Wartenberg (1990), but to date, detection algorithms are rare, aside from those reported in Lu, Chen, and Kou (2004) and Chen et al. (2008), where robust local MDs are used. Fully reporting such advances in this use of a GW methodology is beyond the scope of this study, and is instead presented in Harris et al. (2014). Here, we simply introduce this topic, where only a basic GWPCA-based detection method is investigated. Our use of a nonrobust GWPCA is considered valid, as it should at least detect the most extreme outliers.

We detect outliers with GWPCA simply by investigating the LOOR data that are used to determine an optimal bandwidth (i.e., $\text{LOOR}_i = \sum_{ik} w_i \left([\mathbf{X}]_{ik} - [\mathbf{S}]_{ik}\right)^2$). Thus, when GWPCA is calibrated with an optimal bandwidth using the preceding basic selection procedure, observations can be flagged as outlying that contribute the most to the total GOF statistic for that bandwidth (i.e., the same observations whose influence should be down-weighted when the robust selection procedure is used). However, in the context of multivariate spatial outlier detection, LOOR data sets should be found for a range of user specified bandwidths, covering, say, 5%, 10%, 20%, and 40% of the data. Thus, the determination and nature of an outlier depends on the spatial scale at which it is viewed.

In this study, we choose to present our detection results using GWPCA calibrated with a box-car kernel, using only 7.5% and 100% bandwidths. In the former case, multivariate spatial outliers (termed local outliers) are detected, while in the latter case (effectively a PCA), multivariate outliers are detected (termed global outliers). A box-car kernel is chosen following the preceding discussions. An outlier is an anomalous observation vector (of size m) at a sample location i rather than one particular element in this vector. For each LOOR data set, a cutoff needs to be specified where an observation vector is deemed outlying if it has a LOOR value exceeding this cutoff. Here, the determination of a cutoff that separates background data from anomalies is not straightforward (Filzmoser and Todorov 2013), and with experiments, we investigate four cutoffs, which can be categorized into the following two groups:

(1) A cutoff is taken as the upper whisker of the LOOR data's standard or adjusted box-plot (each specified with defaults; see Rousseeuw et al. 2006). An adjusted box-plot (Hubert and Vandervieren 2008) accounts for skewed data and provides a robust cutoff.

(2) Robust z-scores calculated for the LOOR data are compared with cutoffs set at 2.5 or 3. The LOOR data are robustly standardized by subtracting their median and dividing by their Qn scale estimator (Rousseeuw and Croux 1993). For a variable x, a definition for the Qn scale estimator is $\sigma_{\text{Qn}} = c_1 \cdot c_2 \cdot \{|x_i - x_j; i < j|\}_{(p)}$, where c_1 is a consistency factor depending on data size; $\text{c}_2 = 2.2219$; and $p = \binom{h}{2}$, where $h = [n/2] + 1$. Robust z-scores are suggested in a related context by Daszykowski et al. (2007).

A LOOR data set and its cutoffs depend on the number of components retained in their PCA/GWPCA model, in addition to the chosen bandwidth and kernel.

Supplements: GW correlations and randomization tests

Before embarking on a GWPCA, conducting preliminary local analyses using simpler models, such as GW correlations (Fotheringham, Brunsdon, and Charlton 2002), is useful. Furthermore, for GW correlations and GWPCA, Monte Carlo diagnostic tests are possible for assessing nonstationarity. Tests can confirm whether or not the GW model or aspects of the GW model are significantly different from those found by chance or artifacts of random variation in the data. Here, the sample data are successively randomized, and the GW model is applied after each randomization. A basis of a significance test is then possible by comparing the true result with results from a large number of randomized distributions (ideally 9,999). The randomization hypothesis is that any pattern seen in the data occurs by chance, and therefore any permutation of the data is equally likely.

For GW correlations, at each location i, the test evaluates whether the true GW correlation can be said to be significantly different from such a GW correlation found by chance (e.g., Harris and Brunsdon 2010). For GWPCA, the test evaluates whether the local eigenvalues vary significantly across space (Harris, Brunsdon, and Charlton 2011). Here, the paired coordinates are successively randomized among the variable data set, and after each randomization, GWPCA is applied (with an optimally reestimated bandwidth) and the standard deviation (SD) of a given local eigenvalue is calculated. Next, the true SD of the same local eigenvalue is included in a ranked distribution of SDs. Its position in this ranked distribution relates to whether significant (spatial) variation is present in the chosen local eigenvalue. Unlike a GW correlation test, where the results are mapped, the results from a GWPCA test are graphed.

Case study: freshwater chemistry data for Great Britain

The study data are a subset of those from a water chemistry sampling program for Great Britain that is part of the United Kingdom (U.K.) Department of Transport and Regions freshwater acidification critical loads mapping program (Kreiser, Patrick, and Battarbee 1993). Water chemistry samples were taken during the autumn or early spring over the period 1992–1994. Sites were chosen to represent the most acid-sensitive water body within either a 10-km grid square (for medium to high sensitive areas), or a 20-km grid square (for low or nonsensitive areas), to calculate the minimum critical load. Then research teams within the Critical Loads Advisory Group (CLAG) used the water chemistry data to calculate and map critical load values (for details, see CLAG Freshwaters 1995).

The data subset chosen for our case study is composed of eight water chemistry variables at 533 freshwater sites. The eight variables selected (with units of measure and acronyms in parentheses) are: pH; alkalinity (μeq L^{-1}, Alk); conductivity (μS cm^{-1}, Cond); nitrate or NO_3^- (μeq L^{-1}, NO3); sulfate or SO_4^{2+} (μeq L^{-1}, SO4); phosphate or PO_4 (μeq L^{-1}, PO4); total monomeric aluminum (μg L^{-1}, AL.TM); and total organic carbon (mg L^{-1}, TOC). The full data set consists of 1,335 U.K. freshwater sites and 14 variables. Sites for Northern Ireland and many U.K. islands were removed from the analysis because the use of Euclidean distances in our GW models may not be appropriate with these sites retained. Sites with missing values were also removed. With adaptations for use with different distance metrics and missing data, future work could apply GWPCA to the full data set. Although the focus of the sampling program is freshwater acidification, not all of the 14 variables directly relate to acidification, with some variables also providing information about other environmental concerns, such as freshwater eutrophication. The eight variables selected for our demonstration of GWPCA were expertly considered the most valuable (of the 14) for understanding the nature of freshwater acidification, which is the focus of this case study.

Expected benefits of using a GWPCA

Given that variation in the study data is largely driven by the deposition of various (acidifying and nonacidifying) compounds, types of land use, and geology, all of which show considerable variation, both singly and in combination across Great Britain, the structures and relationships in the study data most likely vary in similar ways across space. Here, a GWPCA allows us to uncover such changes in a single, coherent analysis, which in turn should provide valuable regional insights into acidification and its potentially harmful impacts to freshwater life. Previous PCA studies using subsets of this (or associated) water chemistry data include ones by Bennion, Harriman, and Batterbee (1997) and by Kernan, Hughes, and Helliwell (2002), where PCA was applied to sites in southeast England in the former, while PCA was applied to high-altitude sites in Scotland in the latter. GWPCA provides a means of conducting both studies concurrently (and PCA studies for all regions in between) with a single analysis. Furthermore, we can use GWPCA outputs to partition or classify the water chemistry data into geographical regions, which then can be used to conduct individual (more detailed) PCAs and further statistical analyses because these regions display a certain homogeneity in the structure of their freshwater chemistry data. We may choose to conduct such partitioned analyses only in regions where the water chemistry is highly unusual and/or regions where strongly acidified waters are likely. In this respect, a GWPCA provides a means to target at a national scale, the most susceptible regions to acidification, where remedial actions are most urgent. This approach presents a significant advance over earlier studies where areas were selected because little knowledge of water chemistry relationships existed (e.g., Bennion, Harriman, and Batterbee 1997). As the collection of large environmental data sets continues (e.g., as required for the European Water Framework Directive; http://ec.europa.eu/environment/water/water-framework/), the availability of analysis methods capable of providing a visual overview of complex relationships over substantial geographical regions will become increasingly valuable, largely because it facilitates researchers and regulators identifying target areas where management actions may be required to prevent environmental degradation.

Data preprocessing

All variables other than pH are strongly positively skewed, and as such, seven of the eight variables were jointly transformed to approximate multivariate normality using a multivariate Box–Cox power transform (e.g., Ruppert 2006). Cube-root, fourth-root, and logarithmic transforms were used as convenient approximations to the actual Box–Cox parameters found. Transformed variables are renamed as follows: Alk.T, Cond.T, NO3.T, SO4.T, PO4.T, AL.TM.T, and TOC.T. For the global PCA, we standardize these variables to specify the covariance matrix. The same (globally) transformed and (globally) standardized data are also used in the GWPCA calibrations, which are similarly specified with (local) covariance matrices. For consistency, we use our GWPCA algorithms specified with a box-car kernel and a 100% bandwidth to provide all our PCA outputs.

Preliminary global and local analyses

While some of the study variables may be expected to be strongly correlated (e.g., pH and Alk.T) at both global and local scales, others such as TOC.T and AL.TM.T may exhibit stronger local correlations in areas impacted by acidification, and hence exhibit different global and local correlations. Here, the global linear correlation matrix appears in Table 1, where the strongest correlations are between pH and Alk.T (as expected; $r = 0.92$), and between Cond.T and SO4.T ($r = 0.87$). Weak correlations are also evident, whereas some variables (e.g., NO3.T and AL.TM.T) are uncorrelated. The PCA results (Table 2) reveal that the first three components have eigenvalues greater than unity, and that they collectively account for 79.9% of the variation in the data. This result reflects relatively strong levels of (global) collinearity among some of the variables. From the PCA loadings, component 1 would appear to strongly represent Alk.T and Cond.T; component 2, AL.TM.T and TOC.T; and component 3, NO3.T. These results are partially summarized in the PCA distance biplot (see Jolliffe 2002, pp. 90–107) for the first two components (given in the lower right-hand corner of Fig. 6a), where the variable loadings are represented by directional vectors and the scores data are represented by points labeled with their site ID numbers. Strongly correlated variables tend to have similar directional vectors (e.g.,

Table 1 Global Correlation Matrix

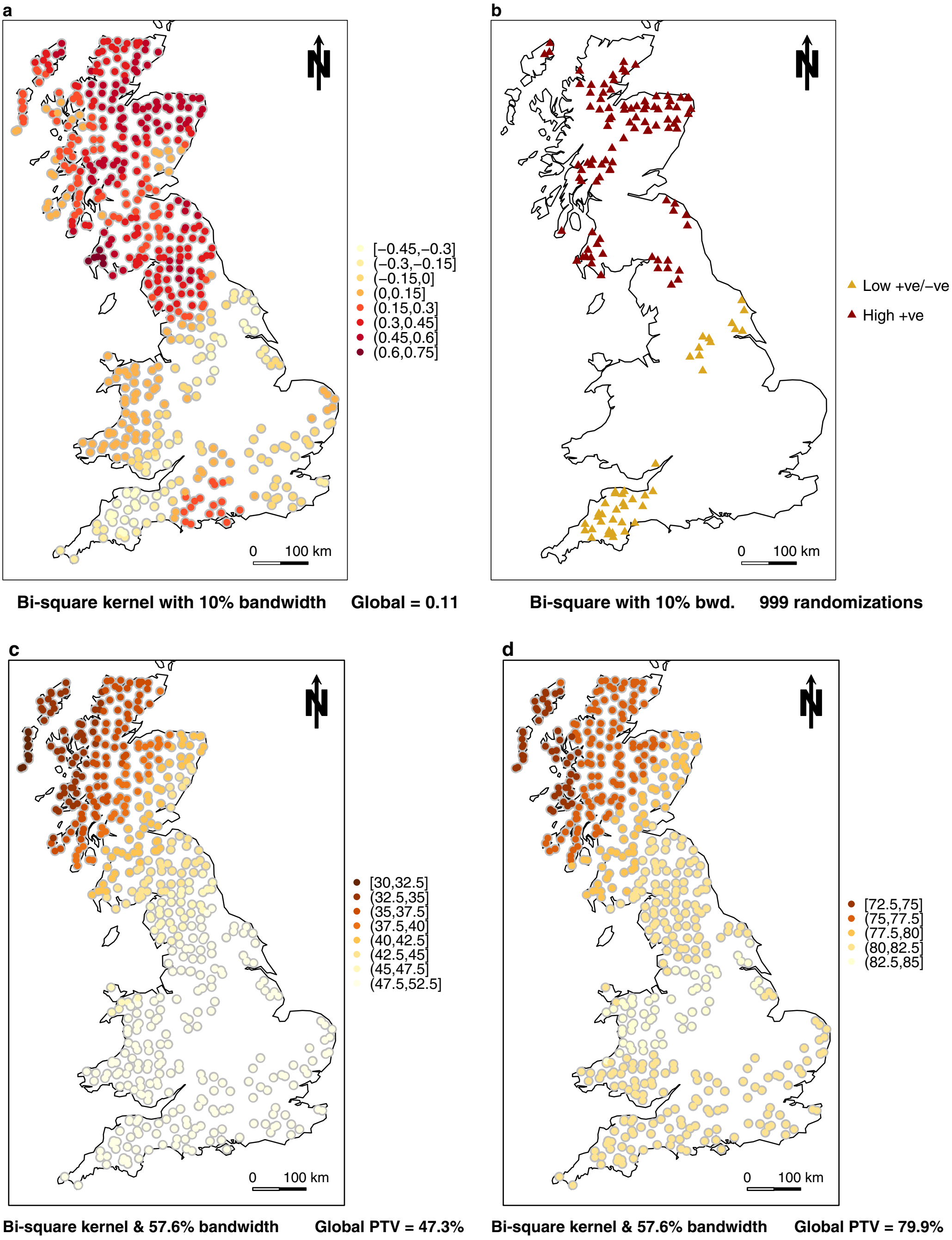
	pH	Alk.T	Cond.T	NO3.T	SO4.T	PO4.T	AL.TM.T	TOC.T
pH	1	0.92	0.58	0.15	0.47	0.19	–0.63	0.10
Alk.T	—	1	0.75	0.21	0.63	0.33	–0.57	0.26
Cond.T	—	—	1	0.30	0.87	0.35	–0.33	0.41
NO3.T	—	—	—	1	0.39	0.18	0.00	–0.06
SO4.T	—	—	—	—	1	0.34	–0.21	0.34
PO4.T	—	—	—	—	—	1	–0.03	0.44
AL.TM.T	—	—	—	—	—	—	1	0.11
TOC.T	—	—	—	—	—	—	—	1

Table 2 Eigenvalues, PTV, Cumulative PTV, and Loadings for the Global PCA

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Eigenvalues	3.775	1.527	1.079	0.691	0.397	0.363	0.110	0.045
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
Loadings:								
pH	0.420	–0.349	–0.106	–0.081	–0.348	0.411	–0.215	–0.590
Alk.T	0.477	–0.182	–0.097	–0.025	–0.211	0.308	0.140	0.756
Cond.T	0.461	0.130	0.063	0.334	0.278	–0.059	0.709	–0.270
NO3.T	0.179	0.140	0.818	–0.246	–0.404	–0.233	0.037	–0.007
SO4.T	0.424	0.199	0.219	0.356	0.431	–0.015	–0.645	0.066
PO4.T	0.240	0.432	–0.169	–0.779	0.321	0.120	0.010	–0.056
AL.TM.T	–0.273	0.543	0.170	0.235	–0.142	0.723	0.063	–0.014
TOC.T	0.202	0.536	–0.449	0.179	–0.534	–0.375	–0.106	–0.026

Cond.T and SO4.T). Fig. 6a is deliberately designed to be viewed and interpreted as a whole; in this respect, any detailed interpretation of an individual biplot is not really viable.

As an example of locally exploring the data with GW correlations, we present site maps for: (i) the GW correlation between TOC.T and AL.TM.T in Fig. 1a; and (ii) a randomization test, at the 95% level of significance, for the same GW correlation in Fig. 1b. The GW correlations are specified using a bi-square kernel with a 10% bandwidth. Our choice of bandwidth is a judged one, chosen so as to provide reasonably reliable local correlations. GW correlations are useful in that they provide insights into the nature of relationship nonstationarity among variable pairs. For our chosen kernel and bandwidth, the correlation between TOC.T and AL.TM.T clearly varies across space from $r = –0.36$ to $r = 0.69$ (compared with a global correlation of $r = 0.11$). The associated randomization test indicates areas of unusually weak negative correlations, mainly in northeast and southwest England. Unusually strong positive correlations are clustered at various locations in Scotland and northern England, areas where acidified waters are likely. Interestingly, however, no locations in Wales are indicated as having unusually strong positive correlations, although this is an area where acidification is known to occur (e.g., Fritz et al. 1989). This outcome may warrant further investigation because acid mine drainage, which results in very high aluminum concentrations (IPCS 1997), may be obscuring the TOC.T/AL.TM.T



Bi-square kernel with 10% bandwidth Global = 0.11

Bi-square with 10% bwd. 999 randomizations

Bi-square kernel & 57.6% bandwidth Global PTV = 47.3%

Bi-square kernel & 57.6% bandwidth Global PTV = 79.9%

Figure 1. (a) GW correlations for TOC.T and AL.TM.T; (b) randomization test results for GW correlations for TOC.T and AL.TM.T; (c) GWPCA PTV for local component 1 (PC1); and (d) GWPCA PTV for the first three local components (PC1 to PC3).

relationship. Similar evidence of correlation nonstationarity is observed for other (but not all) variable pairs, and is taken to support the pursuit of a multivariate local analysis with GWPCA, where relationships for all eight variables can be studied as a whole.

GWPCA bandwidth selection

In the spirit of exploration, we specified 28 GWPCAs resulting from some combination of the following: (i) one-to-seven retained components; (ii) a box-car or bi-square kernel; and (iii) a basic or robust bandwidth selection procedure. Here, for a few specifications, the resultant bandwidth function did not indicate a clear minimum. From the PCA results, proceeding with $q = 3$ components seems natural because it provides a cumulative PTV of 79.9%. As such, we focus our GWPCA investigations on those bandwidth functions where $q = 3$, with the intention of conducting a GWPCA to directly correspond to that of a reasonable PCA specification. These bandwidth functions are presented in Fig. 2, where minimums are reached in all four cases.

With respect to the two different kernel specifications, box-car kernels result in smaller optimal bandwidths than that found with bi-square kernels. This is entirely expected because the distance-decay weighting of the bi-square kernel permits more local information, albeit with decreasing influence from each GWPCA calibration point. With respect to the basic or robust selection procedures, the shapes of the bandwidth functions are similar, with similar minimums. This outcome suggests that outlying observations have minimal influence on bandwidth selection. The presence of outlying observations is indicated by the medians of the LOOR data sets being consistently lower than their respective means. In summary, a GWPCA with three retained components appears appropriate, and we choose a bi-square kernel calibration as our study GWPCA, where an optimal bandwidth is taken at 57.6%.

To provide support for our chosen GWPCA specification, a randomization test was conducted to evaluate whether the local eigenvalues for each component vary significantly across space. The results are given in Fig. 3, where the *P*-values for the true SD of the eigenvalues are calculated at 0.020, 0.939, and 0.657 for the first, second, and third components, respectively. In two of three cases, the null hypothesis of local eigenvalue stationarity is firmly not rejected at the 95% level. However, a GWPCA is still worthwhile because the null hypothesis of local eigenvalue stationarity is firmly rejected at the 95% level for the dominant first component. This, in turn, suggests significant local eigenvalue nonstationarity for the first two and the first three components, combined.

Intuitively, the multivariate structure of the water chemistry data is not expected to remain the same across Great Britain, a feature our investigations reflect. Nonstationary correlations, GWPCA bandwidth functions that reach clear minimums, and associated tests that indicate eigenvalue nonstationarity, all provide compelling evidence to this effect when viewed as a whole. However, finding the exact scale at which a GWPCA should operate is not a straightforward task. Here, our objective analyses have held in check the application of a GWPCA with an inappropriate, subjectively chosen bandwidth, which for this study could have been selected too narrow (e.g., a 10% bandwidth, as that subjectively chosen for the GW correlations). That said, bandwidths can be user specified when there exists some strong prior belief or expert knowledge to do so. The results from an objective analysis can help guide this choice. Similar discussions are often presented in related kernel weighting paradigms (e.g., kernel density estimation), where automated approaches for finding an optimal bandwidth are not viewed as a panacea for bandwidth selection (Silverman 1986).

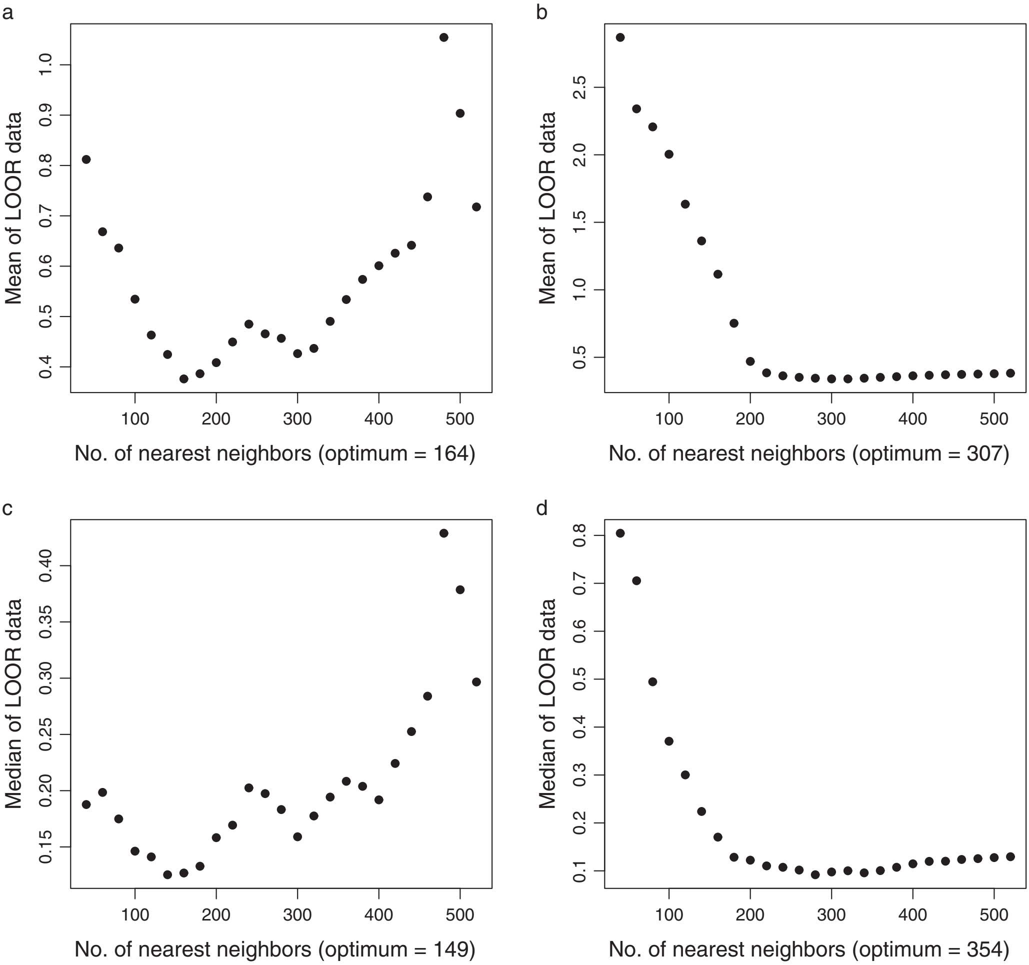


Figure 2. GWPCA bandwidth functions with three retained components: basic procedure with (a) box-car and (b) bi-square kernels; robust procedure with (c) box-car and (d) bi-square kernels. As a percentage, optimums occur at 30.8%, 57.6%, 28.0%, and 66.4%, respectively.

GWPCA visualization

In this section, we present two new ways to visualize the output from a GWPCA, complementing those by Harris, Brunsdon, and Charlton (2011), where maps are presented for: (i) the variable with the highest loading per component; (ii) the loading signs per component; and (iii) the PTV data. As all GWPCA studies should map the PTV data, we present such maps for the first, and the first three components combined, in Fig. 1c and d, respectively. In both maps, the spatial patterns in the PTV data are broadly similar, with higher PTVs generally located in England and Wales, while the lowest PTVs are located in northwest Scotland. It is possible to view the number of components to retain, given some prespecified threshold of total variance preserved. For example, from Fig. 1d, and for a threshold of 80%, only the first three components are required for most of Great Britain (i.e., akin to the global case from Table 1). For much of Scotland, however, the first four (or more) components are required, indicating a regional reduction in

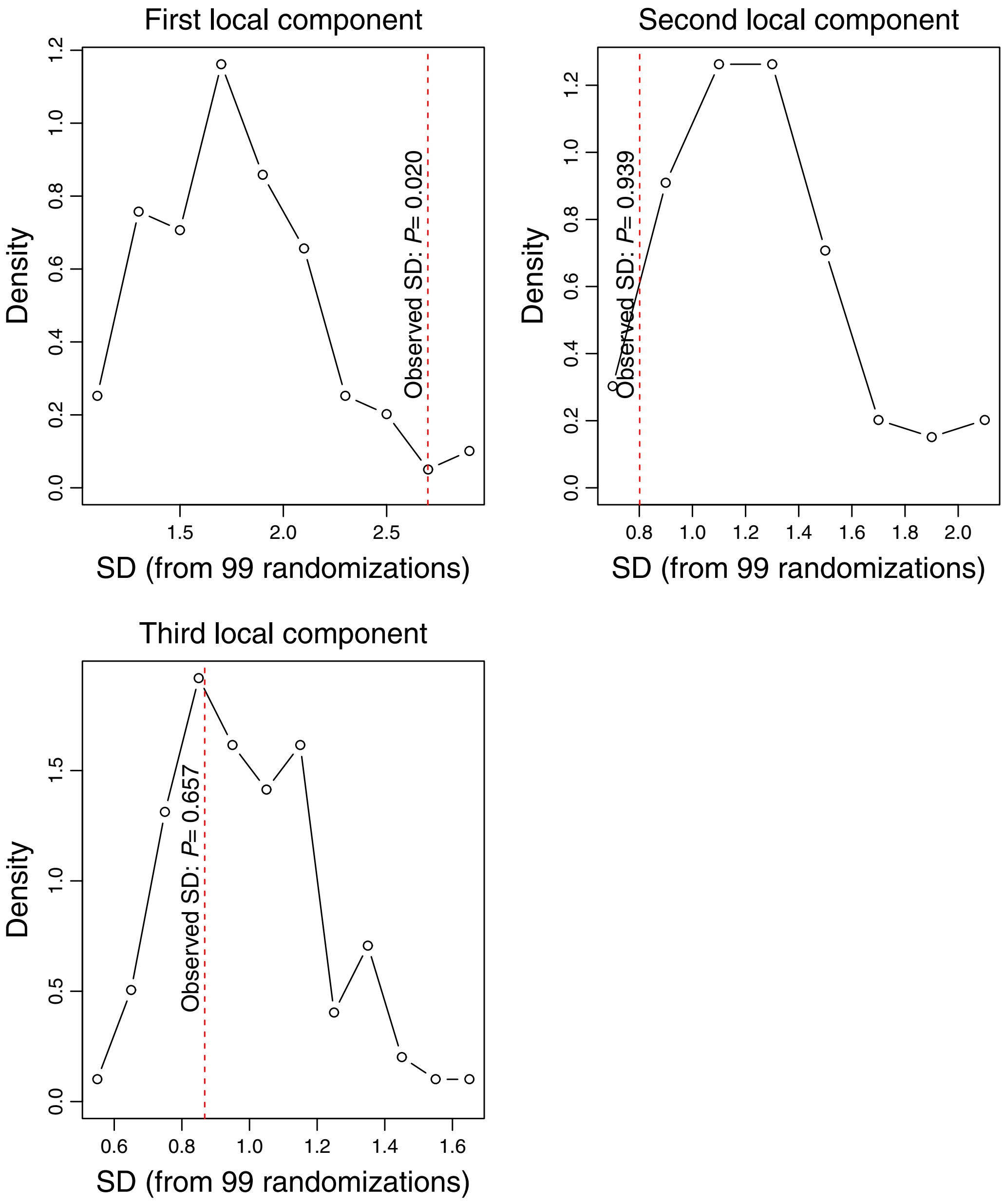


Figure 3. Randomization test for eigenvalue nonstationarity for first, second, and third local components. GWPCA bandwidths were reestimated using the basic procedure.

variable collinearity. Scotland, as a whole, appears to have the most spatially diverse water chemistry data structures. The banded southwest to northeast pattern broadly follows the distribution of major soil types and the underlying geology (e.g., see the Scottish soil transacts maps provided by The Macaulay Land Use Research Institute at http://www.macaulay.ac.uk/tipss/scotst1.htm, last accessed March 3, 2014).

Visualization 1: loading size and sign for all variables per component

A useful visualization is the change in size and sign of the eight local loadings together at each of the 533 freshwater sites. In this respect, we use a multivariate glyph with spokes around a central hub in which the length of a spoke corresponds to the size of its local loading, and its color corresponds to the sign. In this case, gold (light shading) signifies positive and red (dark shading) signifies negative. The glyph is scaled relative to the spoke with the largest absolute loading. The variable corresponding to each local loading is always in the same place on the glyph as follows: pH is at 0° (north); Alk.T is 45° (northeast); Cond.T is 90° (east); NO3.T is 135° (southeast); SO4.T is 180° (south); PO4.T is 225° (southwest); AL.TM.T is 270° (west); and TOC.T is 315° (northeast).

Fig. 4a and b presents two such glyph maps for the first and second components, respectively. Here a spatial preponderance of glyphs of one color or another, or larger spokes on the same variables provide a general indication of the structures being represented at each of the freshwater sites. General trends in the local loadings are visible, with a key trend (for both components) following a north to south direction. Here, similar data structures separately exist in: (i) northern Scotland; (ii) southern Scotland and northern England; and (iii) the rest of England and Wales. However, a visualization problem exists because correctly interpreting the behavior of so many glyphs is difficult.

In order to provide a more interpretable visualization, we adopt a filter approach where the aim is to map only those glyphs that represent the maximum or greatest change in the structure of the water chemistry data. Considering the irregular nature of the study area and the number of spokes on the glyphs (i.e., the number of variables), we aim to reduce 533 site glyphs to eight spatially representative ones. To achieve this, we subject the loadings to a hierarchical clustering procedure for the first and second components via Ward's minimum variance method (Ward 1963), where the resultant dendrogram is cut to yield eight cluster groups. The results (Fig. 4c) appear reasonable, where five of the clusters are found in Scotland, reflecting the known diversity in its water chemistry data structures.

Next, we choose one site from each of the eight cluster groups (Fig. 4c) to act as a group representative. Representative sites are not chosen randomly, but chosen so that the eight resultant site glyphs can be mapped with as large as possible a symbol without overlapping each other. Fig. 5a and b presents the revised (filtered) glyph maps for the first and second components, respectively. These appear with their corresponding glyphs that are found globally with a PCA. Results are much clearer; strong evidence exists of spatial change in the structure of the study data. Here, we can see for the first component that the global glyph only loosely reflects local structure at our chosen site for cluster group 7, representing central Scotland (Fig. 4c). All other local glyphs moderately or strongly differ from the global one. However, none of the local glyphs differ from the global glyph in terms of the loading signs, where loadings for AL.TM.T are consistently negative. For the second component, a greater diversity appears in the water chemistry structure than is found for the first component, where loading size and sign can vary across space. For both components, the general north-to-south trend in the structure of the data has now a certain clarity to it. Overall, the loadings for NO3.T appear to exhibit the greatest geographical variation. For the first component, large loadings in Scotland reflect high variation in NO3.T, whereas small loadings in much of England and Wales indicate lower levels of variation. This outcome is most likely a reflection of agricultural land use being more widespread in the southern part ofGreat Britain, resulting in a higher concentration of nitrate from fertilizers being more generally prevalent.

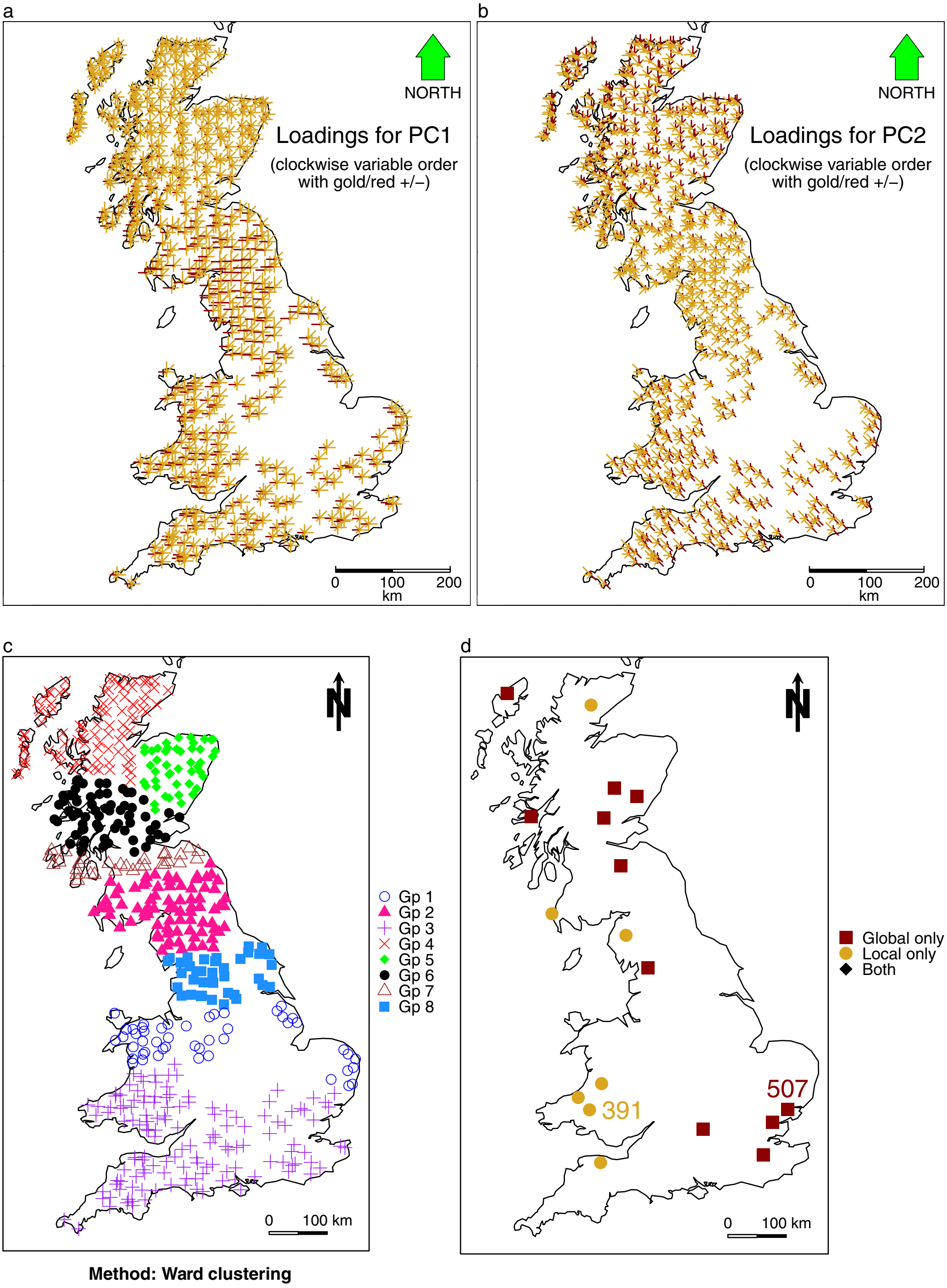


Figure 4. Multivariate glyphs of GWPCA loadings at all locations for (a) local component 1 (PC1) and (b) local component 2 (PC2). Ward cluster analysis specified with eight groups for the GWPCA loadings of PC1 and PC2 given in (c). Location of global and local outliers given in (d)—two sites are highlighted with their site IDs (391 and 507) for further scrutiny (Fig. 7).

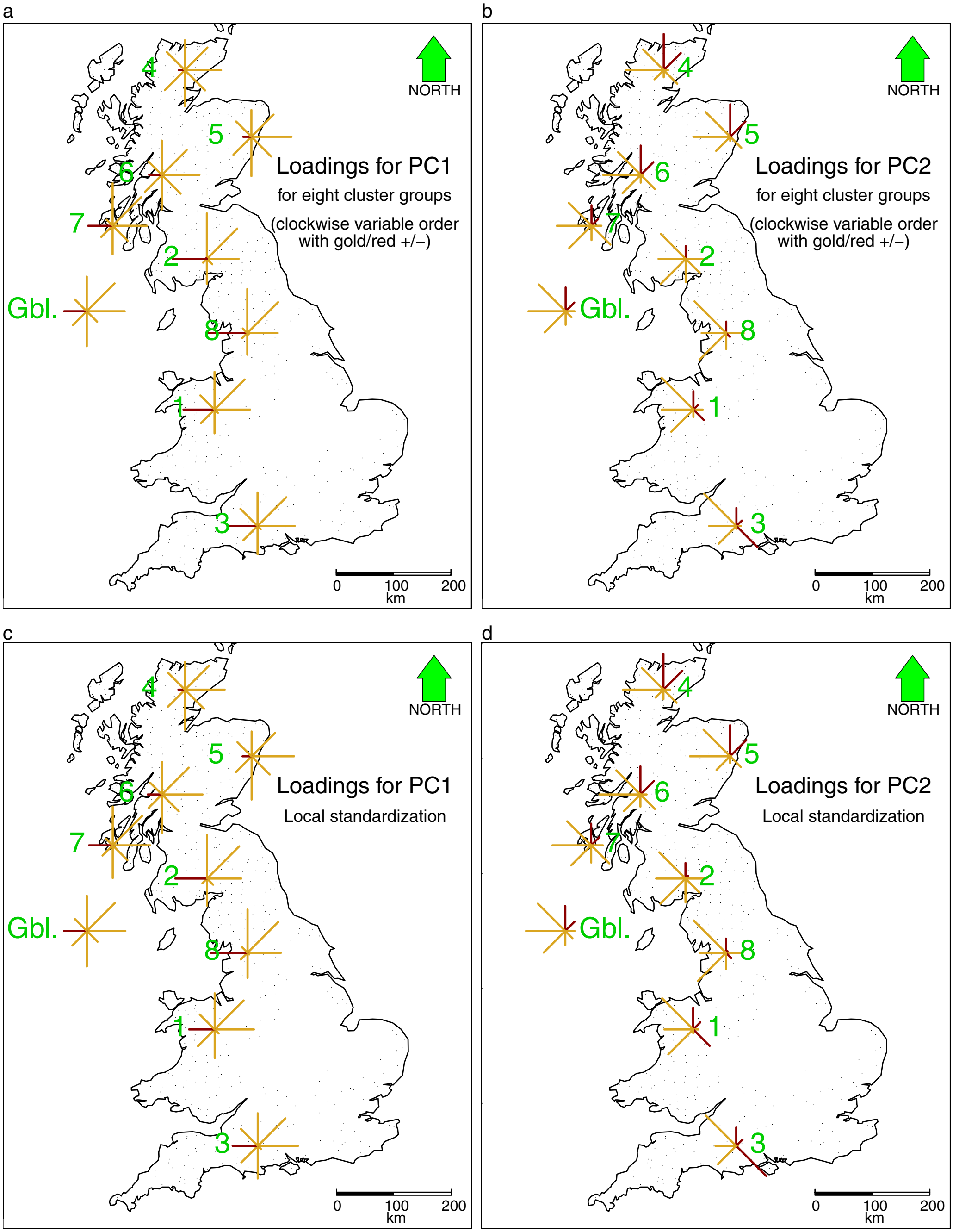


Figure 5. Multivariate glyphs of GWPCA loadings at eight selected locations according to a prior cluster analysis (Fig. 4c) for (a) local component 1 (PC1) and (b) local component 2 (PC2). The same glyphs are given in (c) and (d), but using locally standardized data. Global (Gbl.) glyphs of PCA loadings are also given for PC1 and PC2.

Visualization 2: local biplots for first and second components

Again using the filtered-site approach, we can find a distance biplot for the first two components for each of the same eight locations. These local biplots appear in Fig. 6a, together with the global biplot for context. The local biplots are loosely presented in their geographic positions, with those at the top of the figure representing sites in Scotland. As with the local glyphs, the local biplots should be viewed and interpreted as a whole and in relation to each other. Again, clear evidence exists of spatial change in the structure and the interrelationships of the study data. Focusing on the behavior of the loadings for one variable in relation to the other seven can be informative. For example, NO3.T tends to positions itself with TOC.T and PO4.T in Scotland (at representative sites for five of the cluster groups), while this relationship is not evident in England and Wales. Confirming suspected stationary relationships is also useful. For example, the local relationship between pH and Alk.T remains consistently strong across space, and similar to that found globally.

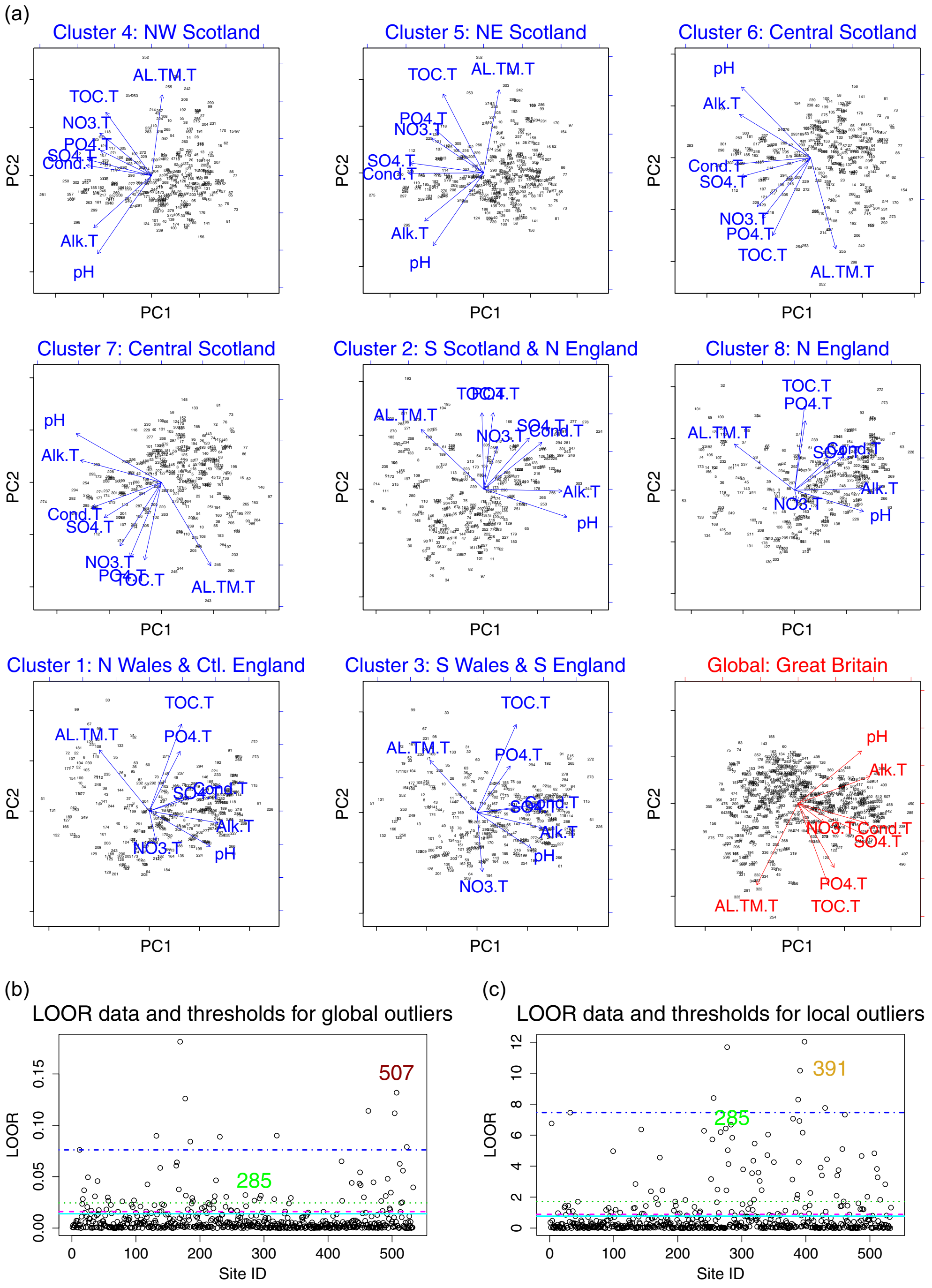
Multivariate spatial outlier detection

Fig. 4d portrays the locations of suspected outlying observation vectors according to the following criterion: (i) a PCA-based method of detection for global outliers; and (ii) the corresponding GWPCA-based method for local outliers. Both methods are specified with $q = 3$ retained components. Experimenting with the different cutoff options for the resultant LOOR data sets allowed the adjusted box-plot procedure to be chosen. Cutoffs are viewed as strict, noting that only the most extreme outliers are likely to be reliably detected with our nonrobust algorithms. Fig. 6b and c plots the LOOR data sets and the competing cutoffs for the global and local detections, respectively. Eleven global and seven local outliers are identified, where no observation vector is both globally and locally outlying. Eight outliers are in Scotland, where their existence may contribute to the diverse data structures observed in that region.

To demonstrate the worth of our GWPCA-based method, Fig. 4d highlights two outlying sites. One site on the east coast of England with an ID No. 507 is a suspected global outlier only. The other site in South Wales with an ID No. 391 is a suspected local outlier only. These sites are good examples of what might happen if only some nonspatial, global method of detection was conducted: (a) a false-positive detection because the spatial neighbors of site 507 are similar in value; and (b) a false-negative detection because the spatial neighbors of site 391 are dissimilar in value.

These misclassifications are depicted using parallel coordinate plots (Inselberg 2002), where we use unweighted (PCPlot) and GW (GWPCPlot) versions. In these plots, each variable is represented by a parallel vertical axis. For the PCPlot, all eight data values are plotted for all 533 sites, where the plot lines are given an equal (transparency) weighting. The line depicting the multivariate structure for the outlying site is colored red (and dashed), while all other lines are colored black. For the GWPCPlot, the line for the outlying site is still colored red (and dashed), while all other lines (in black) are shown with increasing levels of transparency according to their sites' geographic distances from the outlying site. Transparency levels are weighted via a bi-square kernel, where lines for the most distant sites appear fully transparent.

Fig. 7a and b displays the PCPlot and GWPCPlot for site 507. This site is outlying if we only view the data in a nonspatial manner with the PCPlot. However, this site is not outlying if we view the data in a spatial manner with the GWPCPlot, because data at this site are similar in structure to data at nearby sites. Thus, a GWPCA-based method of detection does not consider this site as


(a)

Cluster 4: NW Scotland

Cluster 5: NE Scotland

Cluster 6: Central Scotland

Cluster 7: Central Scotland

Cluster 2: S Scotland & N England

Cluster 8: N England

Cluster 1: N Wales & Ctl. England

Cluster 3: S Wales & S England

Global: Great Britain

(b)

LOOR data and thresholds for global outliers

(c)

LOOR data and thresholds for local outliers

Figure 6. F(a) GWPCA (local) distance biplots for first (PC1) and second (PC2) components at eight selected locations according to a prior cluster analysis (Fig. 4c). PCA (global) distance biplot for PC1 and PC2 is also given. LOOR data and competing thresholds (dotted green line, standard box-plot procedure; dotted/dashed blue line, adjusted box-plot procedure; solid light blue line, robust *z*-score procedure with threshold of 2.5; dashed pink line, robust *z*-score procedure with threshold of 3) are given in (b) for global outlier detection and (c) for local outlier detection.

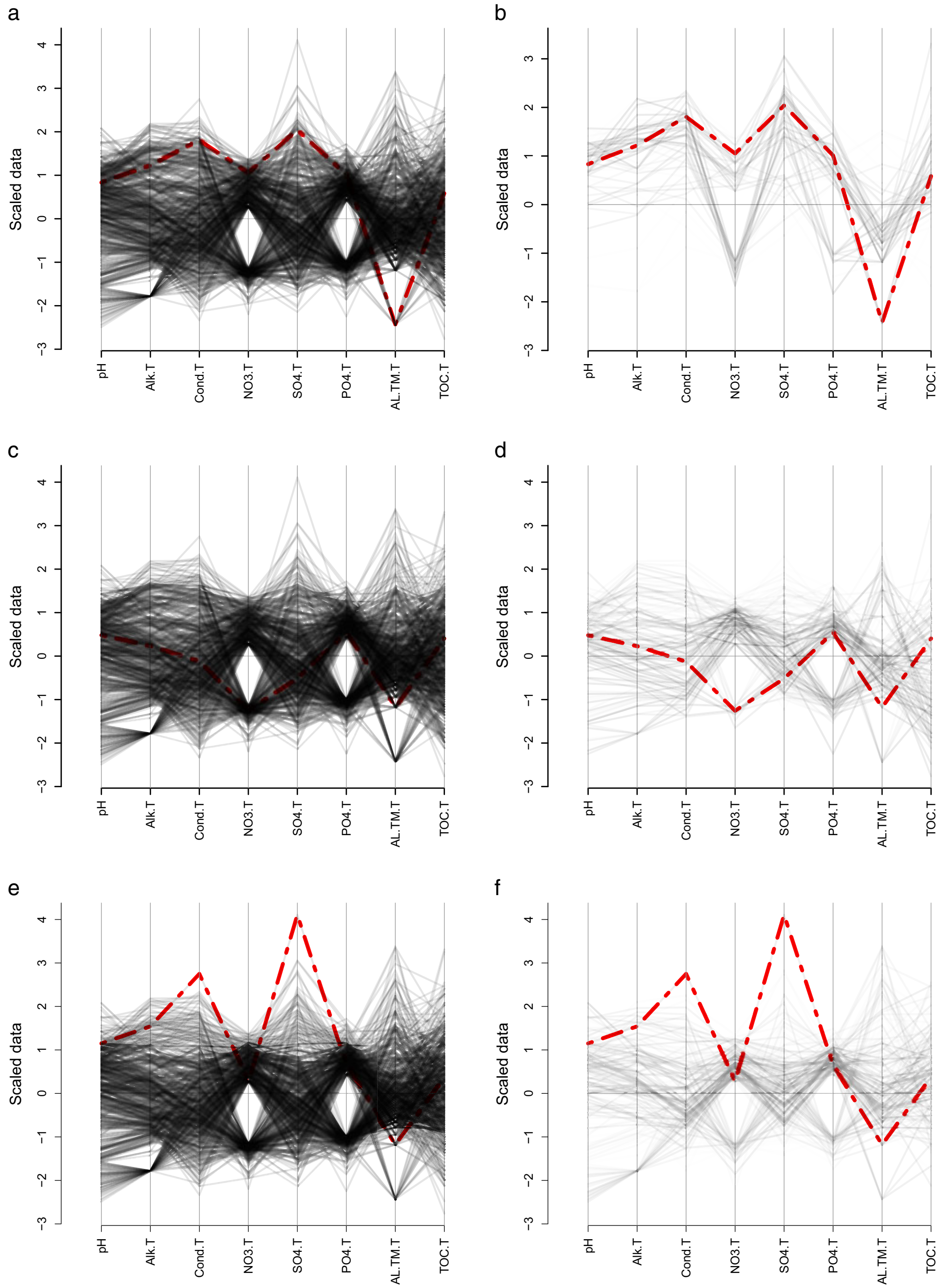


Figure 7. PCPlots for (a) a global outlier (only) at site 507, (c) a local outlier (only) at site 391, and (e) a "global" and "local" outlier at site 285. GWPCPlots for (b) a global outlier (only) at site 507, (d) a local outlier (only) at site 391, and (f) a "global" and "local" outlier at site 285. Outliers are highlighted in a dashed red line. See Figs. 4d and 6b and c.

outlying, while a PCA-based method does. Similarly, Fig. 7c and d displays the plots for site 391. In this case, the site is not outlying if we view the data only with the PCPlot, but is outlying if we view the data with the GWPCPlot. Data at this site tend to be dissimilar in structure to data at nearby sites. Thus, a GWPCA-based method does consider this site as outlying, while a PCA-based method does not. Further plots are given for data at site 285 (located in northern England) in Fig. 7e and f, providing an example of a site that is potentially outlying from both viewpoints. Although this site has not been identified as globally or locally outlying according to our chosen criteria, the site tends to this dual status as suggested by its elevated position in Fig. 6b and c.

Outliers can also be identified via the scores plot of a PCA distance biplot, where outliers relate to points far removed from the majority (i.e., those toward the edge of a biplot). However, these outliers are of a different type than those identified in this study via the LOOR data. Outliers via the scores data are outlying in the PCA subspace, while outliers via the LOOR data are outlying in that they do not fit the subspace model. From the global PCA distance biplot for this study, a number of potentially outlying sites exist with respect to the scores data: sites 285, 496, 497 are associated with maximum values of SO4.T, Cond.T, and NO3.T; and sites 254, 323 are associated with maximum values of AL.TM.T. Aside from site 285, a site appearing as outlying in both senses is unusual. For a fuller discussion on the different types of outliers that can be detected with PCA, see Hubert, Rousseeuw, and Vanden Branden (2005).

Cautionary notes on data preprocessing decisions

Care must be taken when conducting a PCA with raw or transformed data, which then are used in an unstandardized or standardized form, because data preprocessing decisions can strongly affect analytic results (Baxter 1995; Cao, Williams, and Williams 1999). Such decisions are further complicated by the presence of outliers which often are the major cause of any differences observed (Baxter 1995). For this study, we conducted both PCA and GWPCA using transformed data, which then were standardized. As a consequence, all of our study results and interpretations are somewhat dependent on these decisions.

The use of data transforms

Although an assumption of multivariate normality is not required for PCA, this technique does require an assumption of linearity among the variables; thus transforming data should promote linearity (see Varmuza and Filzmoser 2009, pp. 66–67). For PCA and GWPCA, our analysis in the transformed (and standardized) data space was considered to provide clearer and more interpretable output than an analysis in the raw (standardized) data space (which was conducted, but not reported). Fortunately, most results were similar. An analysis with the raw data most likely is compromised by the seven heavily skewed variables, which in part is due to the presence of outlying observations. Transforming the data helps address these issues, as would the adoption of a robust analysis (Harris et al. 2014). For PCA, data transforms can change the structure and relationships in data. GWPCA is similarly affected but now locally, because transforms can also change the data’s spatial structures and correlations. In doing so, this can affect the choice of bandwidth for GWPCA, and thus alter the perception of spatial heterogeneity.

The use of standardized data

Data standardization is required as PCA is not scale invariant. It ensures that the study variables have equal importance via two operations: (i) mean centering and (ii) scaling. The first operation ensures that the component scores are centered at zero. The second operation ensures that variables with the largest variances do not dominate the principal components (as each standardized variable has a variance of 1). Thus, component loadings and scores (commonly) change when a PCA is conducted with standardized data. For this study, the unstandardized data means are 6.70, 3.98, 4.82, −1.77, 5.22, −2.79, 1.94, and 1.64 for pH, Alk.T, Cond.T, NO3.T, SO4.T, PO4.T, AL.TM.T, and TOC.T, respectively. The respective unstandardized data variances are 1.15, 4.99, 0.81, 40.96, 1.24, 41.48, 0.64, and 0.18. As such, the use of unstandardized data (transformed or not) is not considered an option.

Consequences for outlier detection

Similar analytical challenges arise for our PCA/GWPCA-based outlier detection methods, where for this study, we have detected outliers only in a transformed (and standardized) data space. Detection with the raw (standardized) data first, then with the transformed data second, may provide a worthy, albeit more lengthy alternative. Here, it may be useful to remove or truncate the most extreme outliers that are found in the raw data investigation, prior to detections with the transformed data. Knowledge of both raw and transformed data outliers is important, because if subsequent models need to be fitted using transformed data (to promote good fits), then knowledge of outliers in this data space is of value. A data transform can both reduce and increase the number of outliers. For example, in the univariate case, with positively skewed nonzero data and a logarithmic transform, outliers in a 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 Box–Cox transform themselves can be compromised by the existence of outliers (Ruppert 2006).

Should data preprocessing be conducted globally or locally?

A final cautionary note concerns the implications of conducting data preprocessing operations globally and applying GWPCA to these data. If the study data are globally transformed and globally standardized, there is no guarantee that the data will retain their associated properties at the scale of each local PCA. Thus, for GWPCA to conform to the demands of PCA, these data preprocessing operations (at least with respect to standardization) apparently should be conducted locally.

Although both operations are viable locally, currently conducting a full GWPCA using locally preprocessed data is not possible because the component scores would be location specific and incomparable across space. Consequently, the described automatic bandwidth selection procedures would produce meaningless results. Here, each LOOR value (which uses local scores data) would be location specific and incomparable to all other LOOR values, entailing that summed or average LOOR statistics could not be found. The described outlier detection method would be similarly meaningless because it also uses the LOOR data. These problems would still arise when conducting only one of the two local preprocessing operations. Thus, only a limited GWPCA is possible using locally preprocessed data, where the bandwidth must be user specified, and where by assumptions only the PTV data and the component loadings are comparable across space.

In light of these limitations, which defy an obvious solution, data preprocessing for GWPCA should remain global, but complemented with a series of pragmatic data checks. These checks are recommended even when the data do not need to be transformed and/or standardized, because a data set's global characteristics commonly differ from its local characteristics. Ideally, each localized data set of a GWPCA should consist of variables that are roughly normal with means that are close to zero and with variances that are roughly similar to each other; any checks should reflect these features. Checks could proceed formally via multiple hypothesis tests and associated corrections (e.g., Caldas de Castro and Singer 2006). Alternatively, the following informal checks are useful.

A first check is to investigate the local means, variances, skewness, and kurtosis for each variable of the globally preprocessed data, where these statistics are calculated for the same scale of each local PCA of the GWPCA. Thus, for the eight variables of this study, 8 × 4 = 32 local statistic data sets need to be investigated. As an example of this investigation with respect to data standardization only, site 239 of the study data has the most diverse local means, while site 347 has the most diverse local variances. The local means for site 239 are –0.10, –0.18, –0.35, –0.24, –0.34, –0.03, 0.11, and 0.06, whereas the local variances for site 347 are 1.14, 1.16, 0.86, 0.80, 0.72, 0.93, 1.28, and 1.09 for pH, Alk.T, Cond.T, NO3.T, SO4.T, PO4.T, AL.TM.T, and TOC.T, respectively. Although not exactly what is required, a global standardization appears to provide local data sets with means and variances that tend to what is required. Differences simply reflect the underlying spatial variation in each of the eight variables.

A second and complementary check is to conduct a GWPCA with locally preprocessed data, specified with the same bandwidth as that used in the GWPCA with globally preprocessed data. Next, the output of the (limited) GWPCA needs to be compared with that from the original (full) GWPCA. Fig. 5 presents an example of such a comparison for the study data, where we again focus the investigation on data standardization only (i.e., the use of a global transform is retained in both GWPCAs). Fig. 5a and b presents the filtered glyph maps for the first and second components using the original GWPCA, whereas Fig. 5c and d presents the corresponding maps for the alternative GWPCA. These eight sites display little difference in the loadings with respect to global or local data standardizations. Further comparisons were conducted for all 533 sites using both the PTV and loadings data, with the only differences of note occurring in the loadings for southeast England sites.

Discussion and conclusions

In this study, we demonstrate the value of a GWPCA for investigating spatial heterogeneity in the structure of a multivariate environmental data set. Using a freshwater chemistry data set for Great Britain, our chosen GWPCA calibration reveals clear evidence of spatial change in data structures, which otherwise would go unnoticed. As expected, based on the underlying geology and variety of land use present in the region, freshwaters in Scotland appear to have the most spatially diverse water chemistry structures. While most of the findings are in concordance with what is known of general geology/land use/environmental issues and their impact on water quality, a few regional differences are highlighted where atypical processes may be governing water quality. These exploratory findings allow for better informed model decisions for any analysis that may follow. Here, we may opt for a continuous (nonstationary) model, decide on a partitioned approach with separate models, or focus only on those regions where the water chemistry is such that some severe environmental damage is likely. The benefit of a global study of multiple environmental parameters over a large geographical scale should not be ignored at a time when comparable environmental data sets are being compiled both nationally and internationally within the European Union (i.e., the Water Framework Directive). Our approach is considered to have the potential to highlight geographical areas worthy of further investigation that may be difficult to identify with a piecemeal analytical style.

Our study also provides four advancements to the GWPCA methodology. First, a more assured choice in finding the scale at which each local PCA should operate is possible via the use of basic and robust bandwidth selection procedures. Second, improved visualizations are demonstrated, where GWPCA output is mapped only at locations that represent the greatest change in the structure of the study data. To find these representative locations, GWPCA loadings are fed into a clustering algorithm, where the resultant classification map displays reasonable spatially distinct groupings. Related local measures, such as local Moran's *I* (Anselin 1995) combined with other data, often have improved the accuracy of a given spatial classification study (e.g., Emerson, Lam, and Quattrochi 2005; Su et al. 2008; Pant, Singh, and Srivastava 2010), and similar improvements may result if GWPCA loadings are used. GWPCA loadings are useful because they reflect a multivariate local spatial structure for all variables taken together, whereas local Moran's *I* data only reflect a univariate local spatial structure for each variable in turn. Third, we extend GWPCA to detect multivariate spatial outliers. Results are promising and complement similar work using robust detection methods (Harris et al. 2014). A fourth advancement is presented with respect to data preprocessing decisions prior to the application of a GWPCA—that of data transforms and data standardization. Here, a useful guide is given regarding the consequences of these decisions in terms of GWPCA output, when the issue is whether they should be conducted globally or locally.

In conclusion, the idea of geographically weighting can be extended to other unconstrained ordination techniques that are commonly applied in physical geography studies, such as principal coordinate analysis. A more ambitious extension would be to formulate a GW version of some canonical ordination technique, where the aim is to extract and relate the structures of two data sets (e.g., environmental and species abundance data for an application in ecology). Here, a GW redundancy analysis would be a natural starting point because it would combine the well-studied technique of GWR with GWPCA.

Acknowledgements

For Paul Harris and Martin Charlton, research presented in this article was funded by a Strategic Research Cluster grant (07/SRC/I1168) by the Science Foundation Ireland under the National Development Plan. We gratefully acknowledge this support. We also thank the anonymous reviewers and the editor, Dan Griffith, whose comments helped significantly improve this article.

References

Anselin, L. (1995). "Local Indicators of Spatial Association—LISA." *Geographical Analysis* 27, 93–115.

Atkinson, P. M., S. E. German, D. A. Sear, and M. J. Clark. (2003). "Exploring the Relations between River Bank Erosion and Geomorphological Controls Using Geographically Weighted Logistic Regression." *Geographical Analysis* 35, 58–82.

Austin, M. (2007). "Species Distribution Models and Ecological Theory: A Critical Assessment and Some Possible New Approaches." *Ecological Modelling* 200, 1–19.

Baxter, M. J. (1995). "Standardization and Transformation in Principal Component Analysis, with Applications to Archaeometry." *Journal of the Royal Statistical Society* C(44), 513–27.

Bennion, H., R. Harriman, and R. Batterbee. (1997). "A Chemical Survey of Standing Waters in South-East England, with Reference to Acidification and Eutrophication." *Freshwater Forum* 8, 28–44.

Brunsdon, C., A. S. Fotheringham, and M. Charlton. (1996). "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity." *Geographical Analysis* 28, 281–9.

Brunsdon, C., A. S. Fotheringham, and M. Charlton. (1998). "Geographically Weighted Regression—Modelling Spatial Non-Stationarity." *The Statistician* 47, 431–43.

Brunsdon, C., J. MaClatchey, and D. J. Unwin. (2001). "Spatial Variations in the Average Rainfall-Altitude Relationship in Great Britain: an Approach Using Geographically Weighted Regression." *International Journal of Climatology* 21, 455–66.

Caldas de Castro, M., and B. Singer. (2006). "Controlling the False Discovery Rate: A New Application to Account for Multiple and Dependent Tests in Local Statistics of Spatial Association." *Geographical Analysis* 38, 180–208.

Cao, Y., D. D. Williams, and N. E. Williams. (1999). "Data Transformation and Standardization in the Multivariate Analysis of River Water Quality." *Ecological Applications* 9, 669–77.

Chen, D., C. Lu, Y. Kou, and F. Chen. (2008). "On Detecting Spatial Outliers." *Geoinformatica* 12, 455–75.

CLAG Freshwaters. (1995). *Critical Loads of Acid Deposition for United Kingdom Freshwaters, Critical Loads Advisory Group, Sub-Report on Freshwaters*. Penicuik: Institute of Terrestrial Ecology.

Daszykowski, M., K. Kaczmarek, Y. V. Heyden, and B. Walczak. (2007). "Robust Statistics in Data Analysis—A Review Basic Concepts." *Chemometrics and Intelligent Laboratory Systems* 85, 203–19.

Davis, J. C. (2002). *Statistics and Data Analysis in Geology*, 3rd ed. New York: Wiley.

Demšar, U., P. Harris, C. Brunsdon, A. S. Fotheringham, and S. McLoone. (2013). "Principal Components Analysis on Spatial Data: an Overview." *Annals of the Association of American Geographers* 103, 106–28.

Dray, S., P. Legendre, and P. Peres-Neto. (2006). "Spatial Modelling: A Comprehensive Framework for Principal Coordinate Analysis of Neighbour Matrices (PCNM)." *Ecological Modelling* 196, 483–93.

Emerson, C., N. Lam, and D. Quattrochi. (2005). "A Comparison of Local Variance, Fractal Dimension, and Moran's I As Aids to Multi-Spectral Image Classification." *International Journal of Remote Sensing* 26, 1575–88.

Farber, S., and A. Páez. (2007). "A Systematic Investigation of Cross-Validation in GWR Model Estimation: Empirical Analysis and Monte Carlo Simulations." *Journal of Geographical Systems* 9, 371–96.

Filzmoser, P., and V. Todorov. (2013). "Robust Tools for the Imperfect World." *Information Sciences* 245, 4–20.

Filzmoser, P., R. Garrett, and C. Reimann. (2005). "Multivariate Outlier Detection in Exploration Geochemistry." *Computers & Geosciences* 31, 579–87.

Filzmoser, P., R. Maronna, and M. Werner. (2008). "Outlier Identification in High Dimensions." *Computational Statistics and Data Analysis* 52, 1694–711.

Foody, G. M. (2008). "Refining Predictions of Climate Change Impacts on Plant Species Distribution through the Use of Local Statistics." *Ecological Informatics* 3, 228–36.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. (2002). *Geographically Weighted Regression—The Analysis of Spatially Varying Relationships*. Chichester, UK: Wiley.

Fritz, S. C., A. C. Stevenson, S. T. Patrick, P. G. Appleby, F. Oldfield, B. Rippey, J. Natkanski, and R. W. Battarbee. (1989). "Paleolimnological Evidence for the Recent Acidification of Llyn Hir, Dyfed, Wales." *Journal of Paleolimnology* 2, 245–62.

Griffith, D., and C. Amrhein. (1997). *Multivariate Statistical Analysis for Geographers*. Englewood Cliffs, NJ: Prentice Hall.

Griffith, D., and P. Peres-Neto. (2006). "Spatial Modelling in Ecology: the Flexibility of Eigenfunction Spatial Analysis." *Ecology* 87, 2603–13.

Harris, P., and C. Brunsdon. (2010). "Exploring Spatial Variation and Spatial Relationships in A Freshwater Acidification Critical Load Data Set for Great Britain Using Geographically Weighted Summary Statistics." *Computers & Geosciences* 36, 54–70.

Harris, P., A. S. Fotheringham, and S. Juggins. (2010). "Robust Geographically Weighed Regression: A Technique for Quantifying Spatial Relationships Between Freshwater Acidification Critical Loads and Catchment Attributes." *Annals of the Association of American Geographers* 100, 286–306.

Harris, P., C. Brunsdon, and M. Charlton. (2011). "Geographically Weighted Principal Components Analysis." *International Journal of Geographical Information Science* 25, 1717–36.

Harris, P., C. Brunsdon, M. Charlton, S. Juggins, and A. Clarke. (2014). "Multivariate Spatial Outlier Detection Using Robust Geographically Weighted Methods." *Mathematical Geosciences* 46(1), 1–31.

Hornung, M., K. R. Bull, M. Cresser, J. Ullyett, J. R. Hall, S. Langan, P. J. Loveland, and M. J. Wilson. (1995). "The Sensitivity of Surface Waters of Great Britain to Acidification Predicted from Catchment Characteristics." *Environmental Pollution* 87, 207–14.

Hubert, M., and E. Vandervieren. (2008). "An Adjusted Boxplot for Skewed Distributions." *Computational Statistics and Data Analysis* 52, 5186–201.

Hubert, M., P. J. Rousseeuw, and K. Vanden Branden. (2005). "ROBPCA: A New Approach to Robust Principal Component Analysis." *Technometrics* 47, 64–79.

Inselberg, A. (2002). "Visualisation and Data Mining of High-Dimensional Data." *Chemometrics and Intelligent Laboratory Systems* 60, 147–59.

IPCS. (1997). "Aluminium". Environmental Health Criteria No. 194, International Programme on Chemical Safety, World Health Organisation, Geneva. Available at: http://www.inchem.org/pages/ehc.

Jetz, W., C. Rahbak, and J. Lichstein. (2005). "Local and Global Approaches to Spatial Data Analysis in Ecology." *Global Ecology and Biogeography* 14, 97–8.

Johnston, R. J. (1978). *Multivariate Statistical Analysis in Geography*. London: Longman.

Jolliffe, I. T. (2002). *Principal Components Analysis*, 2nd ed. New York: Springer-Verlag.

Jombart, T., S. Devillard, A. Dufour, and D. Pontier. (2008). "Revealing Cryptic Patterns in Genetic Variability by A New Multivariate Method." *Heredity* 101, 92–103.

Jombart, T., S. Dray, and A. Dufour. (2009). "Finding Essential Scales of Spatial Variation in Ecological Data: A Multivariate Approach." *Ecography* 32, 161–8.

Kaspari, M., and S. Yanoviak. (2009). "Biogeochemistry and the Structure of Tropical Brown Food Webs." *Ecology* 90, 3342–51.

Kernan, M., M. Hughes, and R. Helliwell. (2002). "Chemical Variation and Catchment Characteristics in High Altitude Lochs in Scotland, UK." *Water, Air, and Soil Pollution: Focus* 2, 61–73.

Kreiser, A. M., S. T. Patrick, and R. W. Battarbee. (1993). "Critical Loads for UK Freshwaters—Introduction, Sampling Strategy and Use of Maps." In *Critical Loads: Concepts and Applications, ITE Symposium No. 28*, 94–8, edited by M. Hornung and R. A. Skeffington. London: HMSO.

Kumar, S., R. Lal, and C. Lloyd. (2012). "Assessing Spatial Variability in Soil Characteristics with Geographically Weighted Principal Components Analysis." *Computational Geosciences* 16, 827–35.

Legendre, P., and E. Gallagher. (2001). "Ecological Meaningful Transformations for Ordination of Species Data." *Oecologia* 129, 271–80.

Legendre, P., and L. Legendre. (1998). *Numerical Ecology*, 2nd ed. Amsterdam: Elsevier.

Lloyd, C. (2010). "Analysing Population Characteristics Using Geographically Weighted Principal Components Analysis: A Case Study of Northern Ireland in 2001." *Computers, Environment and Urban Systems* 34, 389–99.

Lloyd, C., and I. Shuttleworth. (2005). "Analysing Commuting Using Local Regression Techniques: Scale, Sensitivity, and Geographical Patterning." *Environment and Planning A* 37, 81–103.

Lu, C.-T., D. Chen, and Y. Kou. (2004). "Multivariate Spatial Outlier Detection." *International Journal on Artificial Intelligence Tools* 13, 801–11.

Mather, P. M. (1976). *Computational Methods of Multivariate Analysis in Physical Geography*. London: Wiley.

Miller, J., J. Franklin, and R. Aspinall. (2007). "Incorporating Spatial Dependence in Predictive Vegetation Models." *Ecological Modelling* 202, 225–42.

Páez, A., S. Farber, and D. Wheeler. (2011). "A Simulation-Based Study of Geographically Weighted Regression As A Method for Investigating Spatially Varying Relationships." *Environment and Planning A* 43, 2992–3010.

Pant, T., D. Singh, and T. Srivastava. (2010). "Advanced Fractal Approach for Unsupervised Classification of SAR Images." *Advances in Space Research* 45, 1338–49.

Pretty, J. N., C. F. Mason, D. B. Nedwell, R. E. Hine, S. Leaf, and R. Dils. (2003). "Environmental Costs of Freshwater Eutrophication in England and Wales." *Environmental Science and Technology* 37, 201–8.

Rousseeuw, P. J., and C. Croux. (1993). “Alternatives to Median Absolute Deviation.” *Journal of the American Statistical Association* 88, 1273–83.

Rousseeuw, P. J., M. Debruyne, S. Engelen, and M. Hubert. (2006). “Robustness and Outlier Detection in Chemometrics.” *Critical Reviews in Analytical Chemistry* 36, 221–42.

Ruppert, D. (2006). “Multivariate Transformations.” In *Encyclopedia of Environmetrics*, 2223–2227, edited by A. H. El-Shaarwi and W. W. Piegorsch. New York: Wiley.

Schabenberger, O., and C. Gotway. (2005). *Statistical Methods for Spatial Data Analysis*. London: Chapman & Hall.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Su, W., J. Li, Y. Chen, Z. Liu, J. Zhang, T. Low, I. Suppiah, and S. Hashim. (2008). “Textural and Local Spatial Statistics for the Object-Orientated Classification of Urban Areas Using High Resolution Imagery.” *International Journal of Remote Sensing* 29, 3105–17.

Varmuza, K., and P. Filzmoser. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. Boca Raton, FL: CRC Press.

Ward, J. H. (1963). “Hierarchical Grouping to Optimize an Objective Function.” *Journal of American Statistical Association* 58, 236–44.

Wartenberg, D. (1985). “Multivariate Spatial Correlations: A Method for Exploratory Geographical Analysis.” *Geographical Analysis* 17, 263–83.

Wartenberg, D. (1990). “Exploratory Spatial Analysis: Outliers, Leverage Points, and Influence Functions.” In *Spatial Statistics: Past, Present, and Future*, 133–56, edited by D. Griffith. Ann Arbor, MI: IMAGE.

Wheeler, D., and M. Tiefelsdorf. (2005). “Multicollinearity and Correlation among Local Regression Coefficients in Geographically Weighted Regression.” *Journal of Geographical Systems* 7, 161–87.