Robust geographically weighted principal components analysis and its use in sample re-design

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

[1]National Centre for Geocomputation, National University of Ireland,
Maynooth, Ireland
Tel. +353 1 708 6204, Fax +353 1 708 6456
paul.harris@nuim.ie, http://ncg.nuim.ie/
[2]Geography and Planning, University of Liverpool, UK
[3]School of Geography, Politics and Sociology, University of Newcastle, UK
[4]APEM Ltd, Llantrisant, Wales, UK

KEYWORDS: non-stationarity, location-allocation, freshwater chemistry, acidification

1. Introduction

In this study, we investigate the use of geographically weighted principal components analysis (GWPCA) (Fotheringham et al. 2002) as a means to aid sample re-design for a multivariate spatial data set. In GWPCA, a different localised PCA is computed at each target location**,** and as such, the results vary continuously and smoothly over space allowing them to be mapped and explored. This 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 over-simplistic. Key challenges in GWPCA are: (i) finding the scale at which each localised PCA should operate and (ii) visualising and interpreting the copious output that result from its application (Harris et al. 2011a).

Currently, GWPCA has been used as a means to: (a) explore spatially, multivariate data (e.g. Lloyd 2010; Harris et al. 2011a) or (b) identify multivariate spatial outliers (Harris et al. 2011b; 2012). In the latter studies, robust GWPCA was introduced and we again use robust GWPCA, but now in a sample re-design context so that the influence of any multivariate spatial outliers is reduced. In particular, we perform basic and robust GWPCAs on a freshwater chemistry data set for Great Britain, where the resultant local 'percentage of total variance' (PTV) outputs (for a given number of components) are used as 'demand' data for input into a p-median location-allocation (L-A) algorithm (Teitz and Bart 1968). This two-stage procedure enables the optimal location of a pre-specified number of freshwater sites that should provide maximum spatial and multivariate water chemistry information for a re-designed (i.e. second) sampling campaign, where information given by the components is measured by PTV. Justification for this choice of measurement follows in the next section.

Our approach is general and could optimally locate any (pre-specified) number of sample sites to any location (sampled or un-sampled in the first campaign). It could also be applied to other measures of information. However for this particular study, only sites that were visited in the first campaign can be chosen and it is assumed that a much reduced second sampling campaign is required due to limited resources.

2. Methodology

For a given sample re-design, we postulate that less freshwater sites should be selected in areas where the local covariance amongst the water chemistry data is strong, reflected by high local PTV outputs from the GWPCA. In these areas, the local correlation (or local collinearity) amongst the water chemistry data is assumed high, suggesting: (i) less sites are needed and (ii) for sites that are selected, not all variables need to be sampled. Conversely, more sites should be selected in areas where the local covariance amongst the data is weak, reflected by low local PTV GWPCA outputs. In these areas, the local correlation amongst the water chemistry data is assumed low, suggesting: (a) more sites are needed and (b) all variables carry important information and need to be sampled.

Following this postulation, our approach proceeds as follows:

1. perform a (basic or robust) GWPCA on the water chemistry data where the bandwidth that controls the geographical weighting is chosen according to the methodology of Harris et al. (2011a). In this study, bandwidths for an adaptive bi-square kernel are found;

2. by experiment, choose an appropriate number of components to retain and map the PTV outputs. In this study, two components are retained;

3. find the 'residual' PTV data using the formula 100-PTV and use this data as the 'demand' data for input into the L-A algorithm;

4. find the distance matrix for the study data and run the L-A algorithm with a pre-specified number of sites to optimally locate. In this study, 25 sites are pre-specified;

5. map the results from the L-A algorithm indicating the level of 'demand' at each of the chosen sites. The L-A algorithm also 'allocates' nearby sites to each chosen site. This 'allocation' could be mapped to provide insight into the results with respect to the initial sampling campaign, but is not strictly necessary in this sample re-design context.

Robust GWPCA (RGWPCA) is performed in addition to basic GWPCA (BGWPCA), to assess the extent to which BGWPCA output is distorted by outliers. Outliers can strongly influence BGWPCA by artificially increasing local variation and masking key features of the local data structures. For RGWPCA, we estimate the GW covariance matrices using the minimum covariance determinant (MCD) estimator (see Maronna et al. 2006).

3. Case study data and global analyses

The case study data is composed of eight water chemistry variables at 533 freshwater sites across Great Britain. The data is a subset of data used in a freshwater acidification critical loads mapping programme (CLAG Freshwaters 1995). The variables are: pH, alkalinity (Alk.T), conductivity (Cond.T), nitrate (NO3.T), sulphate (SO4.T), phosphate (PO4.T), total monomeric aluminium (AL.TM.T) and total organic carbon (TOC.T). All variables aside from pH were transformed to approximate normality using a power transform. We also standardised the data and specified any PCA with the covariance matrix. The same (globally) standardised data was also used in the GWPCA calibrations, which were similarly specified with (local) covariance matrices. Table 1 summarises the results for the basic and robust PCAs. Basic results reveal that the first three components collectively account for 81.5% of the variation in the data. Component one would appear to strongly represent alkalinity and conductivity; component two, aluminium and total organic carbon; and component three, nitrate. Robust results are different, but not greatly so.

Table 1. Cumulative PTV and loadings from basic and robust PCA.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
BASIC								
Cumulative PTV:	49.6	68.8	81.5	89.5	94.0	97.9	99.3	100
Loadings:								
pH	-0.397	0.376	0.146	0.047	-0.451	-0.351	0.275	0.525
Alk.T	-0.468	0.147	0.122	0.018	-0.198	-0.304	-0.234	-0.746
Cond.T	-0.454	-0.106	-0.063	-0.332	0.289	0.001	-0.665	0.377
NO3.T	-0.258	-0.080	-0.759	0.286	-0.371	0.361	-0.046	-0.006
SO4.T	-0.418	-0.166	-0.233	-0.410	0.388	-0.021	0.642	-0.114
PO4.T	-0.271	-0.419	0.159	0.750	0.331	-0.195	0.064	0.104
AL.TM.T	0.250	-0.567	-0.266	-0.211	-0.288	-0.643	-0.055	0.038
TOC.T	-0.199	-0.543	0.484	-0.174	-0.440	0.450	0.076	0.003
ROBUST								
Cumulative PTV:	45.5	68.6	81.3	89.2	94.4	98.0	99.5	100
Loadings:								
pH	-0.467	0.348	0.290	-0.071	0.511	-0.088	0.046	-0.549
Alk.T	-0.424	0.119	0.201	-0.047	0.281	-0.094	-0.074	0.819
Cond.T	-0.419	-0.152	-0.038	0.437	-0.315	-0.191	-0.674	-0.136
NO3.T	-0.387	-0.182	-0.660	-0.312	0.157	0.501	-0.091	-0.035
SO4.T	-0.409	-0.202	-0.215	0.372	-0.183	-0.254	0.713	-0.011
PO4.T	-0.188	-0.304	0.127	-0.739	-0.302	-0.461	-0.013	-0.077
AL.TM.T	0.239	-0.569	-0.168	0.127	0.641	-0.388	-0.114	-0.019
TOC.T	-0.126	-0.594	0.590	0.062	-0.009	0.519	0.087	-0.050

4. Local analyses and sample re-design

For BGWPCA, an optimal adaptive bandwidth is found whose radius extends to 60.2% of all observed data locations (i.e. the nearest 321 neighbours are weighted), which is associated with the retention of four components. RGWPCA suggested a larger bandwidth at 86.3% (the nearest 460 neighbours) but for comparison, we specify our RGWPCA with the same bandwidth as that used in BGWPCA.

For BGWPCA, the spatial distribution of PTV for the first two components combined is given in Figure 1a. There is clear spatial variation in the BGWPCA results, where smaller PTVs tend to occur in the local case than in the global case. Highest PTVs are located in SW England and Wales, whilst lowest PTVs are located in NW Scotland. For RGWPCA, the corresponding PTV map is given in Figure 1b. Here a different spatial pattern emerges to that found with BGWPCA. Now larger PTVs always occur in the local case than in the global case. Furthermore, highest PTVs are now located in two distinct areas of SE and N England, whilst lowest PTVs are located in two distinct areas of central and S Scotland/N England. Regardless of the GWPCA specification, Scotland appears to have the most spatially-diverse water chemistry data structures.

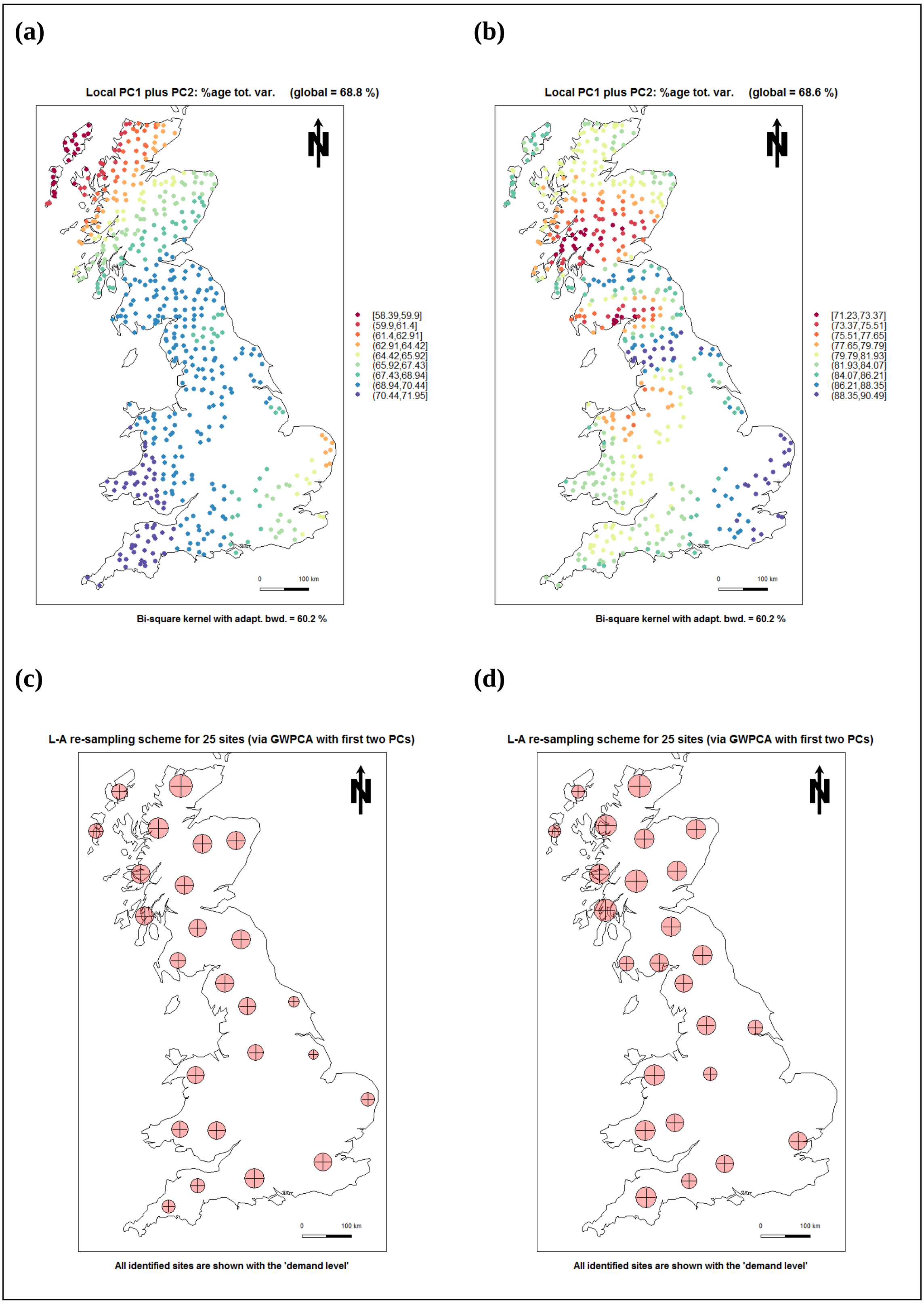
(a)
Local PC1 plus PC2: %age tot. var.   (global = 68.8 %)
N
[58.39,59.9]
(59.9,61.4]
(61.4,62.91]
(62.91,64.42]
(64.42,65.92]
(65.92,67.43]
(67.43,68.94]
(68.94,70.44]
(70.44,71.95]
0
100 km
Bi-square kernel with adapt. bwd. = 60.2 %
(b)
Local PC1 plus PC2: %age tot. var.   (global = 68.6 %)
N
[71.23,73.37]
(73.37,75.51]
(75.51,77.65]
(77.65,79.79]
(79.79,81.93]
(81.93,84.07]
(84.07,86.21]
(86.21,88.35]
(88.35,90.49]
0
100 km
Bi-square kernel with adapt. bwd. = 60.2 %
(c)
L-A re-sampling scheme for 25 sites (via GWPCA with first two PCs)
N
0
100 km
All identified sites are shown with the 'demand level'
(d)
L-A re-sampling scheme for 25 sites (via GWPCA with first two PCs)
N
0
100 km
All identified sites are shown with the 'demand level'

Figure 1. **(a)** Basic and **(b)** robust GWPCA PTV data for the first two components. Sample re-design using L-A with **(c)** basic and **(d)** robust GWPCA outputs as inputs. 'Demand level' is reflected by the size of the pink circle at chosen sites.

Using the L-A algorithm calibrated with the basic and robust GWPCA PTV data, the optimal location of 25 sites for re-sampling is given in Figures 1c and 1d, respectively. As would be expected from the observed differences in the PTV data, the location of sites for a second sampling campaign depends on which GWPCA specification is preferred. For the robust sample re-design procedure, key differences with the basic procedure are noted for: (i) the Hebrides off NW Scotland, where the same two sites are chosen but 'demand' is reduced (i.e. smaller pink circles); (ii) Scotland as a whole, where more sites are chosen with a certain clustering in central and southern regions (including the border area with N England); and (iii) the eastern coast of England, where significantly fewer sites are required. In all such areas, it can be assumed that the existence of multivariate spatial outliers has compromised the basic procedure.

5. Concluding remarks

In summary, an L-A algorithm calibrated with GWPCA outputs provides a simple means to design a second sampling campaign for multivariate spatial data. Future work will consider further experiments and tests, using different data sets. Furthermore, it is possible to weight the 'demand' data. In this study, such a weighting could reflect freshwater sites of most concern to ecological damage (e.g. critical loads for acidification could be used). Future work will consider such weightings. Our approach can be considered a multivariate extension of the univariate approach of Kanaroglou et al. (2005) used to optimally locate air pollution monitors. In this respect, it is applicable to any pollution study, be it water, soil or air pollution, where a high number of contaminants are routinely measured.

Acknowledgements

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

References

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

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

Harris P, Brunsdon C, Charlton M (2011a) Geographically weighted principal components analysis. *International Journal of Geographical Information Science* 25:1717-1736

Harris P, Brunsdon C, Charlton M (2011b) Multivariate spatial outlier detection using geographically weighted principal components analysis. *7th International Symposium on Spatial Data Quality,* Coimbra, Portugal

Harris P, Brunsdon C, Charlton M (2012) Multivariate spatial outlier detection: a comparison of techniques. *geoENV 2012*, Valencia, Spain

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

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* 34:389-399

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

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

Biography

Paul Harris is a Postdoctoral Research Fellow at the National University of Ireland at Maynooth with research interests in spatial statistics with applications in environmental pollution.

Chris Brunsdon is Professor of Human Geography at the University of Liverpool with research interests in spatial data analysis and geocomputation with applications in crime and health.

Martin Charlton is a Senior Research Associate and director of the StratAG program at the National University of Ireland at Maynooth, with research interests in all areas of geographical information science.

Steve Juggins is a Senior Lecturer in Physical Geography at the University of Newcastle upon Tyne, with research interests in palaeolimnology and recent environmental change, diatom-based environmental monitoring, analysis of ecological and palaeoecology data, environmental databases.

Annemarie Clarke is a Senior Consultant Aquatic Ecologist with APEM Ltd. She is an expert on diatom taxonomy and is currently working on water quality and drought monitoring projects.