International Journal of Geographical Information Science
Vol. 25, No. 10, October 2011, 1717–1736

Taylor & Francis
Taylor & Francis Group

Geographically weighted principal components analysis

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

[a]*National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland;* [b]*Department of Geography, University of Leicester, Leicester, UK*

(*Received 26 November 2010; final version received 9 January 2011*)

Principal components analysis (PCA) is a widely used technique in the social and physical sciences. However in spatial applications, standard PCA is frequently applied without any adaptation that accounts for important spatial effects. Such a naive application can be problematic as such effects often provide a more complete understanding of a given process. In this respect, standard PCA can be (a) replaced with a geographically weighted PCA (GWPCA), when we want to account for a certain spatial heterogeneity; (b) adapted to account for spatial autocorrelation in the spatial process; or (c) adapted with a specification that represents a mixture of both (a) and (b). In this article, we focus on implementation issues concerning the calibration, testing, interpretation and visualisation of the location-specific principal components from GWPCA. Here we initially consider the basics of (global) principal components, then consider the development of a locally weighted PCA (for the exploration of local subsets in attribute-space) and finally GWPCA. As an illustration of the use of GWPCA (with respect to the implementation issues we investigate), we apply this technique to a study of social structure in Greater Dublin, Ireland.

Keywords: PCA; GWPCA; bandwidth selection; visualisation; nonstationarity; GWR

1. Introduction

Principal components analysis (PCA) is a widely used technique in the social and physical sciences. Originally developed by Pearson (1901), the details of extracting components for a data matrix and their interpretation were presented in Hotelling (1933). Of the many uses and applications of PCA, the following list of eight by Jeffers (1967) illustrates some of the more common ones:

(1) examination of the correlations between variables of a selected set;
(2) elimination of variables that contribute relatively little information;
(3) examination of the grouping of individuals in *n*-dimensional space;
(4) determination of the weighting of variables in the construction of indices;
(5) allocation of individuals to previously demarcated groups;
(6) recognition of misidentified individuals;
(7) orthogonalisation of regression calculations; and
(8) reduction of the basic dimensions of variability in the measured set.

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

ISSN 1365-8816 print/ISSN 1362-3087 online

http://dx.doi.org/10.1080/13658816.2011.554838
http://www.tandfonline.com

With respect to (7), principal components provide variables that do not exhibit collinearity, and variable-specific regression parameter estimates may be obtained from the component-specific estimates and the inverse of the associated loadings (Joliffe 1982). With respect to (8), the analyst may retain components accounting for, perhaps, 75% of the original variance or only those components with an eigenvalue greater than unity. In many areas of the social sciences, some interpretation of the components is frequently attempted as they can be thought of as new variables or indices, whose character reflects those of the constituent variables with the highest loadings. This usually begins with the component with the largest eigenvalue and is not attempted for any component with an eigenvalue less than unity. Component scores have been used in demographic classification systems (Charlton *et al.* 1985) or as the basis for measures of deprivation (Kelly and Teljeur 2007).

In this study, we are concerned with spatial applications where a standard PCA is replaced with a spatial variant. For spatial problems, there are a host of applications; and according to Gould (1967) these include

(a) measures of terrain roughness;
(b) the varying spatial nature in the connectivity of towns;
(c) orientations of physical features and transport networks;
(d) characteristics of mean information fields (Hägerstrand 1967);
(e) classification;
(f) homogeneity of architectural features;
(g) measures of residential desirability; and
(h) the interpretation of mental maps.

Applications of PCA to ever more diverse geographical problems is somewhat typified by the recent work of Reades *et al.* (2009), where the space-time structure of a mobile phone network is modelled.

Many applications ignore any spatial characteristics in the data and simply apply a standard (aspatial) PCA. However, such effects are often vital to a more complete understanding of a given process and here PCA can be replaced with geographically weighted PCA (GWPCA) (Fotheringham *et al.* 2002, pp. 196–202) when we want to account for a certain spatial heterogeneity. Alternatively, PCA can be adapted to account for spatial autocorrelation in the spatial process (Jombart *et al.* 2008); and a natural extension would be to adapt this spatial PCA technique locally to provide a local spatial autocorrelation GWPCA hybrid. In this article, we focus on the former variant, where we examine some problems concerning the specification and interpretation of a GWPCA. We first consider the basics of (global or standard) principal components, the development of locally weighted PCA (LWPCA) in the exploration of local subsets in attribute-space and finally GWPCA in the exploration of local subsets in geographic-space. As an illustration of the use and interpretation of GWPCA, we use as a case study a data set containing some indicators of social structure in Greater Dublin, Ireland.

The only known application of GWPCA beyond the original work of Fotheringham *et al.* (2002) is provided by Lloyd's (2010) study of population characteristics in Northern Ireland, where a call is made for more research into the visualisation of the output of GWPCA, so that the merits of the technique can be more fully realised. We examine the ways of addressing this and other aspects of the GWPCA methodology and, in doing so, provide an introduction to a more extensive body of work on this subject. In particular, we investigate GWPCA issues of (1) calibration, (2) testing, (3) interpretation and (4) visualisation. In addition, we consider the potential use of GWPCA with respect to an

investigation of collinearity in the geographically weighted regression (GWR) model. All study algorithms were developed within the R statistical computing environment (Ihaka and Gentleman 1996).

2. Methods: from PCA to LWPCA to GWPCA

2.1. Principal components analysis

Given a data matrix $\mathbf{X}$ with n rows representing the observations and m columns representing the variables, the variance–covariance matrix $\mathbf{\Sigma}$ is $m \times m$ with the variances in the leading diagonal and the covariances in the off-diagonal elements. The trace of $\mathbf{\Sigma}$ is the total variance in the data. If the columns in $\mathbf{X}$ are standardised with zero mean and unit variance, the values in $\mathbf{\Sigma}$ will be those in the correlation matrix for $\mathbf{X}$ with its trace equivalent to the number of columns in $\mathbf{X}$. A standard result in linear algebra states that

$$\mathbf{LVL}^{\mathrm{T}} = \mathbf{R} \tag{1}$$

where $\mathbf{V}$ is a diagonal matrix of eigenvalues, $\mathbf{L}$ is a matrix of eigenvectors and the matrix $\mathbf{R}$ is symmetric and positive definite. If $\mathbf{R}$ is a correlation or variance–covariance matrix denoted by $\mathbf{\Sigma}$, then the eigenvalues in $\mathbf{V}$ will represent the variances of the corresponding principal components. The eigenvectors in $\mathbf{L}$ are column vectors and represent the loadings of each variable on the corresponding principal component. It is usual to report the results for the components in decreasing order of eigenvalue; the component with the largest eigenvalue is that which has the largest variance. On dividing by tr($\mathbf{V}$), the eigenvalues can be reported as the proportion of variance accounted for by the corresponding components. Component scores are found by post-multiplying the original data values $\mathbf{X}$ by $\mathbf{L}$; the correlation matrix for $\mathbf{XL}$ is an identity matrix. Component scores are thus a linear combination of the original data values, and given the values of the scores and the loadings, the original data values can be recovered by an inverse transformation.

We can conceive of the data matrix and its eigen-decomposition as representing a *global* model of the covariance structure of the matrix. In spatial settings, the covariance structure of the data is assumed to be constant across the spatial extent of the study area. Thus, although the component scores may be mapped, the eigenvectors and their associated eigenvalues are spatially stationary or *whole-map statistics* in Openshaw *et al.*'s (1987) terminology.

2.2. Locally weighted PCA

LWPCA provides a natural progression towards GWPCA as they are both similar in design. The similarity is analogous to the association between locally weighted regression in attribute-space (Cleveland 1979) and GWR (Brunsdon *et al.* 1996). Furthermore, just as locally weighted regression can be used for geographical problems when the coordinate data are used as attributes (or covariates), LWPCA, similarly calibrated with coordinate data, also has numerous applications. For example, LWPCA can be used to detect edges and other geometric features in LiDAR point clouds, where the weighting is applied directly to the three-dimensional coordinate data.

For LWPCA, the homogeneity of the covariance (or correlation) structure is assumed for those observations that are close to one another in attribute-space. Due to the effect of

inter-variable correlation on the orthogonality of the raw data, inter-observation distances may be measured using Mahalanobis distance D_{M}, as follows:

$$D_{\mathrm{M}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_1 - \mathbf{x}_2)} \tag{2}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are multivariate vectors for two observations and $\boldsymbol{\Sigma}$ is the variance–covariance matrix. A set of (local) principal components can be extracted for each observation in $\mathbf{X}$ using either (i) the nearest observations within some Mahalanobis distance τ or (ii) the nearest k observations, as the (local) calibration data set. Here the bandwidth (or controlling parameter), τ or k, is commonly referred to as fixed (by distance) or adaptive (varying distance), respectively. In either form, the bandwidth may be (a) supplied exogenously by the analyst or (b) estimated from the data through a minimisation of the size of the scores of those components corresponding to lower eigenvalues, using either a leave-one-out or a computationally simpler holdback (set-aside) method (analogous to the GWPCA approach described in Section 4.1). Regardless of the form of bandwidth specified, the local eigen-structure is

$$\mathbf{L}_i \mathbf{V}_i \mathbf{L}_i^{\mathrm{T}} = \boldsymbol{\Sigma}_i \tag{3}$$

with respect to the local subregion of the ith observation, and there will be n sets of eigenvectors and their associated eigenvalues. The scores for the ith observation on the m variables are those for the ith row in the matrix $\mathbf{X}_i \mathbf{L}_i$.

We have described LWPCA in its simplest form, where only a box-car kernel weighting function is specified (for an adaptive bandwidth, the attribute-space weights w_{ij} accord to $w_{ij} = 1$ if $D_{\mathrm{M}_{ij}} \leq \tau_v$ and $w_{ij} = 0$ if $D_{\mathrm{M}_{ij}} > \tau_v$, where $D_{\mathrm{M}_{ij}}$ is the Mahalanobis distance between observations at i and j; and where the Mahalanobis distance τ_v varies accordingly to the bandwidth k). A more generalised description of LWPCA is possible when any kernel function is specified, such as the distance-decay function specified with GWPCA in Section 2.3. Furthermore, LWPCA has the useful property that the eigenvalues and eigenvectors can be estimated at unobserved data locations and not just the observed (sample) locations. As with GWPCA, LWPCA defaults to the global PCA model, if a suitably large bandwidth is specified (i.e. if $k = n$ in this simple case).

2.3. Geographically weighted PCA

At some scale an assumption of multivariate normality is required for PCA, LWPCA and GWPCA. For PCA this operational scale is global; for LWPCA it is local in attribute-space; and for GWPCA it is local in geographic-space. Thus for GWPCA, a vector of observed variables $\mathbf{x}_i$ at spatial location i is assumed to have a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance–covariance matrix $\boldsymbol{\Sigma}$, that is, $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Furthermore, if spatial location i has coordinates (u, v), then PCA with local geographical effects involves regarding $\mathbf{x}_i$ as conditional on u and v, and making $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ functions of u and v; thus $\mathbf{x}_i \,|(u, v) \sim N(\boldsymbol{\mu}(u, v), \boldsymbol{\Sigma}(u, v))$. As $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are functions of u and v, this implies that each element of $\boldsymbol{\mu}(u, v)$ and $\boldsymbol{\Sigma}(u, v)$ is also a function of u and v. Therefore the moments $\boldsymbol{\mu}(u, v)$ and $\boldsymbol{\Sigma}(u, v)$ are the geographically weighted (GW) mean vector and the GW variance–covariance matrix, respectively. To obtain the GW principal components, the decomposition of the GW variance–covariance matrix provides the GW eigenvalues and GW eigenvectors. The product of the ith row of the data matrix with the GW

eigenvectors for the *i*th location provides the *i*th row of GW component scores. The GW variance–covariance matrix is

$$\mathbf{\Sigma}(u, v) = \mathbf{X}^{\mathrm{T}}\mathbf{W}(u, v)\mathbf{X} \tag{4}$$

where $\mathbf{W}(u, v)$ is a diagonal matrix of geographic weights that can be generated using some kernel function. In the case study, we use a bi-square function:

$$w_{ij} = (1 - (d_{ij}/r)^2)^2 \text{ if } d_{ij} \leq r\,; \quad w_{ij} = 0 \text{ otherwise} \tag{5}$$

where the bandwidth is the geographic distance r and d_{ij} is the distance between spatial locations of the *i*th and *j*th rows in the data matrix $\mathbf{x}$. As with any GW model, other kernel shapes are also possible. The GW principal components for the location (u_i, v_i) can be written as

$$\mathbf{LVL}^{\mathrm{T}}|(u_i, v_i) = \mathbf{\Sigma}(u_i, v_i) \tag{6}$$

where $\mathbf{\Sigma}(u_i, v_i)$ is the GW variance–covariance matrix for location (u_i, v_i). In the remainder of this article we refer to GWPCA as *local* PCA.

3. Case study: Greater Dublin social structure

The data set used in this empirical example is composed of eight variables that measure characteristics of social structure in the 322 electoral divisions (EDs) forming Greater Dublin, Ireland. These variables have previously been used in a study of voter turnout (VT) in the Irish 2004 Dáil elections (Kavanagh 2006) in which the results from a GWR analysis (with VT the dependent variable) suggested that there are distinct spatial variations in the social structure of the population of Greater Dublin. GWPCA is used to further uncover and interpret this spatial variation with respect to the independent variables of the same regression study. The variables of interest are the percentage of the population in each ED who are

(a) one-year migrants (i.e. moved to a different address 1 year ago) (Diff.Add);
(b) local authority renters (LA.Rent);
(c) social class one (high social class) (SC.1);
(d) unemployed (Unemp);
(e) without any formal education (Low.Edu);
(f) age group 18–24 (AGE.18.24);
(g) age group 25–44 (AGE.25.44); and
(h) age group 45–64 (AGE.45.64).

Observe that none of the case study variables constitute a closed system (i.e. the full array of values sum to 100) and as such, we do not need to transform the data before the calibration of our PCA models; Lloyd (2010) presented an instance when this is necessary.

As is common practice, we standardise the data and specify the global PCA with the covariance matrix. The effect of this is to make the eight study variables have equal importance (Chatfield and Collins 1980, p. 62). The same (globally) standardised data are also used in the GWPCA calibrations, which are similarly specified with (local) covariance matrices. Further work will investigate the use unstandardised data, or the use of locally

standardised data with GWPCA. Results will almost certainly differ, as all PCA techniques are not scale invariant. For the data of this study, although measured on the same scale, variables are not of a similar magnitude.

The global PCA (Table 1) reveals that the first three components have eigenvalues greater than or very close to unity and that they collectively account for 73.6% of the variation in the data. Component one would appear to represent older residents, component two affluent residents and component three young working residents with lower educational attainment. These are whole-map statistics and interpretations. Whilst representing a Dublin-wide average, they may not represent local social structure particularly reliably. A map of the study area is given in Figure 1 depicting the scores on the first component, where their spatial trend tends to confirm that larger values of the component scores are

Table 1. Summary of global PCA.

	PC-1	PC-2	PC-3	PC-4	PC-5	PC-6	PC-7	PC-8
Eigenvalues	2.878	2.041	0.951	0.840	0.550	0.293	0.248	0.175
Percentage of total variation	36.08	25.59	11.92	10.53	6.89	3.68	3.11	2.20
Loadings								
Diff.Add	−0.389	−0.444	−0.004	−0.149	0.123	0.293	0.445	0.575
LA.Rent	−0.441	0.226	0.144	0.172	0.612	0.149	−0.539	0.132
SC.1	0.130	−0.576	−0.030	−0.135	0.590	−0.343	0.076	−0.401
Unemp	−0.361	0.462	0.022	0.189	0.197	−0.085	0.670	−0.355
Low.Edu	−0.131	0.308	−0.362	−0.861	0.079	−0.062	−0.065	−0.011
AGE.18.24	−0.237	−0.080	0.845	−0.359	−0.224	−0.051	−0.045	−0.200
AGE.25.44	−0.436	−0.302	−0.317	0.053	−0.291	0.448	−0.177	−0.546
AGE.45.64	0.493	0.118	0.179	−0.144	0.289	0.748	0.142	−0.164

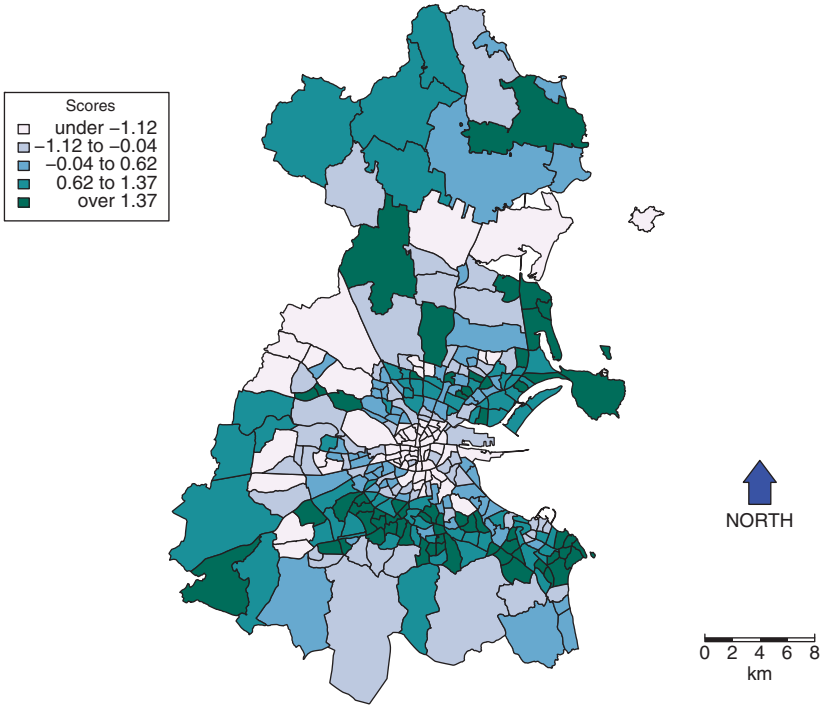

Figure 1. Study area (Greater Dublin) and component one scores from global PCA.

Table 2. Global correlation matrix (Correlations below −0.5 and above +0.5 in bold).

	Diff.Add	LA.Rent	SC.1	Unemp	Low.Edu	AGE.18.24	AGE.25.44	AGE.45.64
Diff.Add	**1**	0.28	0.37	0.01	−0.03	0.34	**0.70**	**−0.56**
LA.Rent		**1**	−0.29	**0.67**	0.17	0.25	0.31	−0.46
SC.1			**1**	**−0.59**	−0.27	−0.03	0.09	0.09
Unemp				**1**	0.28	0.11	0.13	−0.37
Low.Edu					**1**	0.00	0.03	−0.07
AGE.18.24						**1**	0.13	−0.21
AGE.25.44							**1**	**−0.69**
AGE.45.64								**1**

associated with greater proportions of older residents. The global correlation matrix is also given (Table 2). High levels of collinearity amongst the eight variables are evident and this knowledge can help in our interpretation of the PCA (and GWPCA) results.

4. GWPCA: issues and uses

GWPCA brings a set of interesting opportunities and poses a number of challenges. The m variables of a PCA yield m components, m eigenvalues, m sets of component loadings and m sets of component scores. Furthermore, it is conventional to consider a subset of the components with larger eigenvalues as these contribute to the greatest part of data variance. The component loadings can be subject to some interpretation as a composite variable, although this can be difficult. Furthermore, the component scores can be mapped to gain some insights into the spatial distribution of each composite variable (i.e. see Figure 1). For GWPCA, these investigations and interpretations all take place locally; that is, there are m components, m eigenvalues, m sets of component loadings and m sets of component scores at each data location in the study area. Furthermore, we can also obtain eigenvalues and their associated eigenvectors at unobserved locations, although as no data exist for these locations, we cannot obtain scores. This permits the generation of spatial surfaces of eigenvalues and eigenvector loadings.

4.1. Calibration: automatic bandwidth selection

A major challenge in GWPCA (and LWPCA) is in the estimation of the bandwidth. If there are m variables in the data matrix, so that each observation is a vector in m-dimensional space, the scores corresponding to components $q + 1$ to m from PCA represent the Euclidean distances along the axes of the corresponding orthogonal vectors to a q-dimensional linear sub-space. In PCA, the q-dimensional sub-space is spanned by the first q loadings (viewed as m-dimensional vectors) and is the sub-space that maximised the variance of the data points projected on to that sub-space. Here, q is commonly chosen so that this sub-space contains a reasonably high proportion of the total variance, and thus components $q + 1$ to m represent the deviation from this sub-space.

Suppose that $\mathbf{M}_q$ denotes the matrix $\mathbf{M}$ with all but the first q columns removed and $\mathbf{M}_{(-q)}$ denotes the matrix $\mathbf{M}$ with the first q columns removed. Based on Section 2, the first q components are described by $\mathbf{XL}_q$ and the remaining components by $\mathbf{XL}_{(-q)}$. It is possible to show 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 from this $\mathbf{S}$, given by $\mathbf{S} = \mathbf{X} - \mathbf{XL}_q\mathbf{L}_q^{\mathrm{T}}$ can also be written as

$\mathbf{S} = \mathbf{X}\mathbf{L}_{(-q)}\mathbf{L}_{(-q)}^{\mathrm{T}}$ (Joliffe 2002). In effect, through principal components, we find the minimum of the expression $\sum_{ij}([\mathbf{X}]_{ij} - [\mathbf{S}]_{ij})^2$ with respect to $\mathbf{S}$ where $\mathbf{S}$ is a rank q matrix. The problem is solved with the expression above. The variance levels of the components of the matrix $\mathbf{S}$ therefore measure the 'goodness of fit' (GOF) of the projected sub-planes and as such:

$$\mathrm{GOF_i} = \sum_{j=q+1}^{j=m} s_{ij}^2 \tag{7}$$

is the GOF for the ith observation and s_{ij} is the jth component score for observation i, that is, the ijth 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 ith location represent a similar projection, but with the corresponding loadings defined locally. That is, in this case we find $\mathbf{S}$ to minimise $\sum_{ij} w_i([\mathbf{X}]_{ij} - [\mathbf{S}]_{ij})^2$ where w_i is a locally defined weight for location i. The GOF statistic is defined in an analogous fashion as for global PCA; with the exception that in each locality, $\boldsymbol{S}$ is defined using local weights, as above. The GOF statistic provides the means of finding an optimal bandwidth for GWPCA by using either a leave-one-out method or a holdback sample when computing the terms of the statistic.

Using the study data, we demonstrate the calibration of a GWPCA using both fixed and adaptive bi-square kernels, where optimal bandwidths are found using a leave-one-out method. The difficulty now lies in that, unlike the global PCA case, we need to decide *a priori* upon the number of components to retain (i.e. the value of q). Furthermore, we cannot find an optimal bandwidth if we wish to retain all eight components and in this case the bandwidth would need to be user specified. Thus in the spirit of exploration, we find the GOF results with $q = 1, \ldots, 7$ and the two different kernel forms. The resultant bandwidth functions are given in Figure 2, where in many cases an optimal bandwidth is not clearly defined; this is especially so for the fixed bandwidth kernels. Furthermore, there is no pattern to indicate the number of components that should be retained (and there is no reason why there should be one). This behaviour may be in part due to the particular properties of our chosen data set. However, we would recommend a similarly detailed calibration study in all applications of GWPCA.

For our data, a GWPCA with three retained components ($q = 3$) specified with an adaptive bandwidth of $k = 223$ observations would be a natural starting point if we wanted to compare a GWPCA with the global PCA discussed in Section 3(that similarly retains three components). The bandwidth for this particular case suggests a broad but significant spatial nonstationarity in the components (i.e. clear variation in social structure), where approximately two thirds of the study data are used to calibrate each local PCA. If we wish to conduct a GWPCA using all eight components, then all of the results of this investigation can guide the choice of a user-specified bandwidth. As an optimal bandwidth (fixed or adaptive) is fairly clearly defined for seven retained components, then this may be a reasonable starting point.

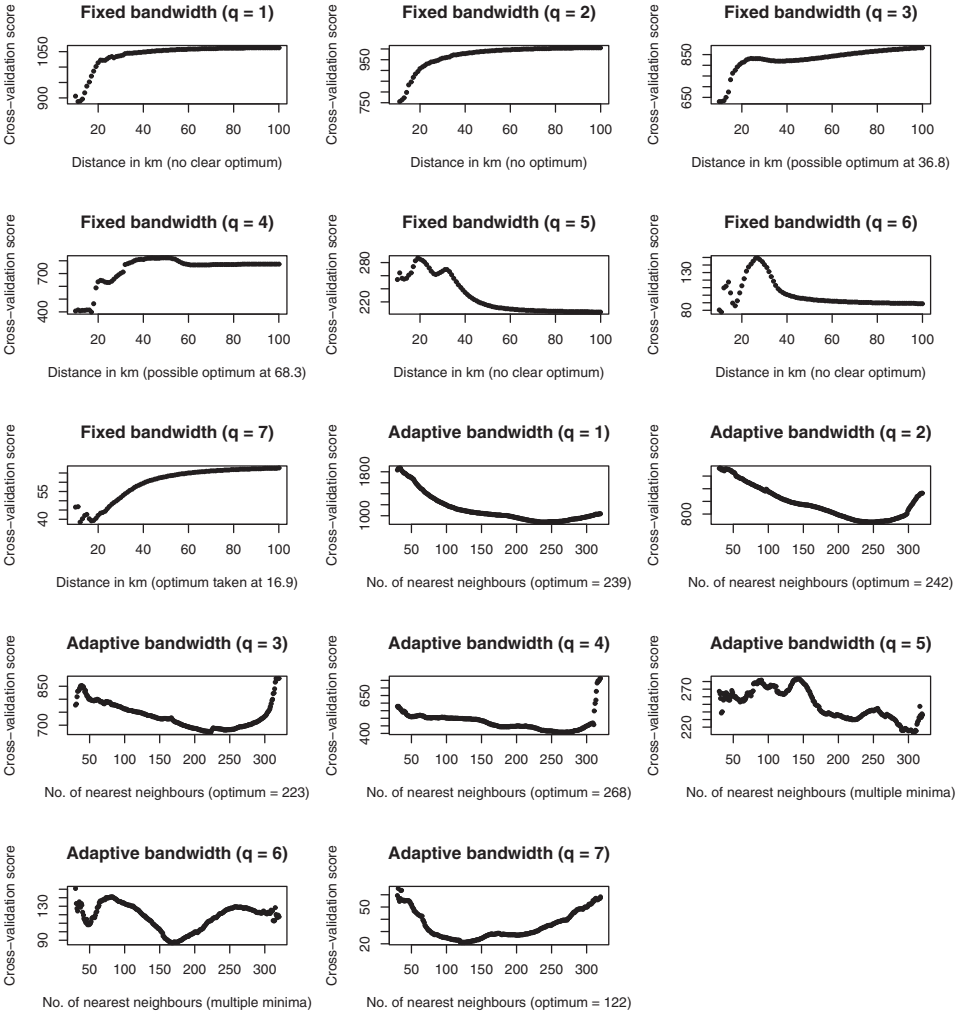

Figure 2. Fixed and adaptive bandwidth functions for different values of q (number of retained components). The study area is approximately 28 km east to west and 44 km north to south.

4.2. Randomisation tests for significance of eigenvalue variability

Monte Carlo tests can be used to evaluate whether local eigenvalues from GWPCA vary significantly across space and so justify the use of GWPCA in the first place. Here the (paired) sample locations are successively randomised amongst the variable data set and after each randomisation, GWPCA is applied and the standard deviation (SD) of a given local eigenvalue is calculated. The actual or true SD of the same local eigenvalue is then included in a ranked distribution of SDs. Its position in this ranked distribution relates to whether there is significant (spatial) variation in the chosen local eigenvalue. A similar procedure has been used in the GWR model to test whether each of its local regression parameters vary significantly across space (Brunsdon *et al*. 1998).

As an example of this test, the resultant distribution of local eigenvalue SDs for the first eigenvalue for the study data is given in Figure 3a. The adaptive bandwidth is re-estimated

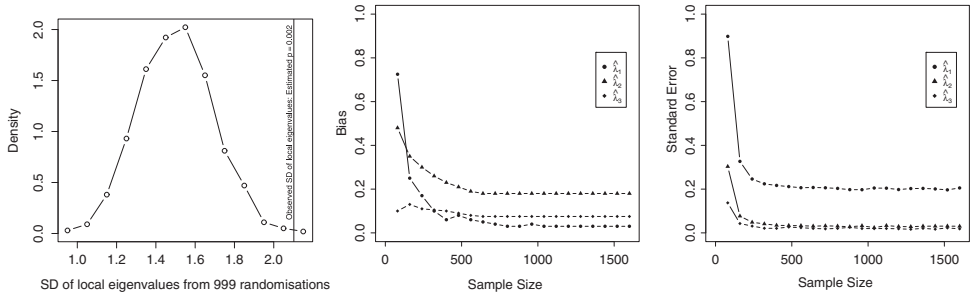

Figure 3. (a) Randomisation test for eigenvalue nonstationarity (case study data); (b) bias and (c) standard error of eigenvalue estimates with increasing sample size (simulation study data).

for each randomised data set and in this case, three components are retained in the GWPCA for bandwidth estimation using the leave-one-out method. As the p-value for the true SD is very small ($p = 0.002$), then the null hypothesis of local eigenvalue stationarity is rejected. This test can be simplified by simulating always using the same (user-specified) bandwidth. This less realistic test would be appropriate if we were interested in a GWPCA that retains all (eight) components.

4.3. Interpretation challenges: experiments with simulated data

It is important to have some handle on the extent that GWPCA can recover local spatial structures in a given data set. In this respect, we briefly report on experiments in which a known spatial structure is imposed on the component eigenvectors and eigenvalues, for some simulated data with a multivariate Gaussian distribution. The results demonstrate that GWPCA is able to recover the known structures reliably, which gives some confidence that the local eigen-structures we observe are not the result of random variation. A set of spatially varying eigenvalues was defined by

$$\text{Eigenvalue}(\mathbf{u}) = \left(u_1^2 + 0.1, (1-u_2)^2 + 0.1, \frac{1}{2}|u_1 - u_2| + 0.1\right) \tag{9}$$

and the corresponding eigenvectors by

$$\text{Eigenvector}(\mathbf{u}) = (u_1\mathbf{I} + (1-u_2)\mathbf{Q})^{\perp} \tag{10}$$

where $\mathbf{u}$ is a location in space; $\mathbf{M}^{\perp}$ denotes the application of the Gram–Schmidt orthonormalisation procedure to any matrix $\mathbf{M}$; and $\mathbf{Q}$ is

$$\mathbf{Q} = \begin{bmatrix} 0 & \frac{4}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{bmatrix} \tag{11}$$

The Gram–Schmidt procedure ensures that the eigenvectors are orthogonal. In this case, several data sets with increasing size n were generated and the estimates of the local eigenvalues λ_1, λ_2 and λ_3 at location $\mathbf{u} = 0$ were generated. The plots in Figure 3b and c show

how the bias and standard error of the estimates vary with sample size. Each of these quantities tends to 0, suggesting that as sample size increases, the mean squared error of the estimates (equal to the sum of the squares of bias and standard error) also tends to 0.

4.4. *Visualisation*

For visualising the output from GWPCA, we may begin by mapping the spatial distribution of the local eigenvalues; there will be m maps, one for each component. The largest value possible for an eigenvalue will be m and it will always be on the first component, so it may help to map the percentage of the total variance that each component accounts for, rather than the raw value. For the study data, Figures 4 and 5 show the spatial distribution of the first and the first three local components combined, respectively. Now all eight components ($q = 8$) need to be retained in our GWPCA model and as such, a user-specified bandwidth (taken at $k = 122$ observations) is also necessary. In both maps, there is clear spatial variation in the local PCA results. A higher percentage of the total variance is generally accounted for in the first component in the local case than in the global case (Figure 4). This is also a characteristic of the first three components combined (Figure 5). The spatial patterns in both plots are similar – much of the local variance is accounted for by the first local component in the centre of Dublin and the surrounding northern and

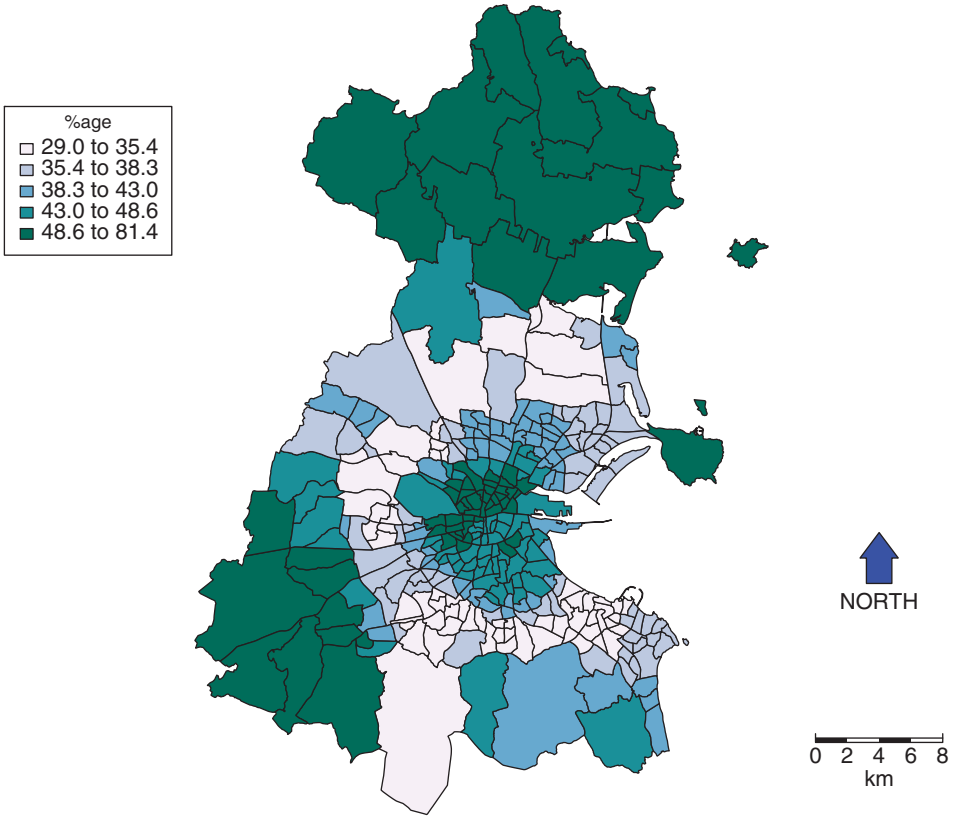

Figure 4. GWPCA ($k = 122$ and $q = 8$) output: percentage of total variation for local component one (at each of 322 EDs). Globally with PCA, this percentage is 36.1%.

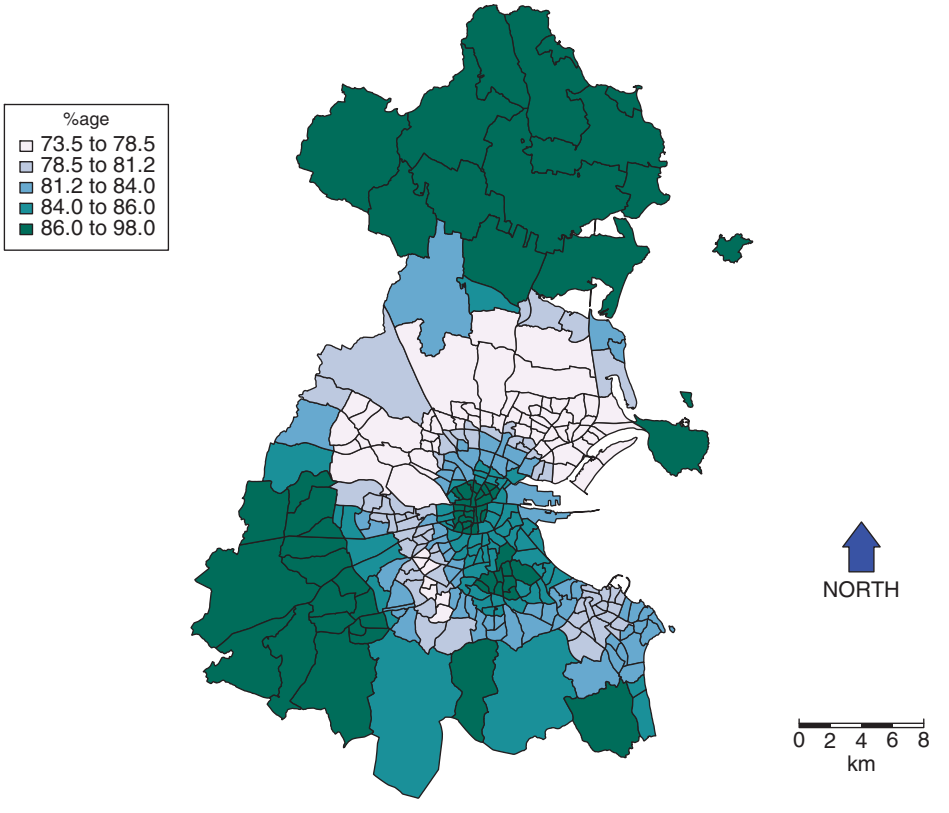

Figure 5. GWPCA ($k = 122$ and $q = 8$) output: percentage of total variation for first three local components (at each of 322 EDs). Globally with PCA, this percentage is 73.6%.

south-western extremities. The inner suburbs, by contrast, would appear to be more diverse in their structures.

In exploring the output from this particular GWPCA model further, we can also map the 'winning' variable in each local component (i.e. the one with the highest absolute local loading). This visualisation is suggested in Fotheringham *et al.* (2002) and the 'winning' variable for the first component is shown in Figure 6. Comparing this with the pattern in Figure 4, the local-authority renting variable appears to play an important part in defining the local structure in central Dublin and educational attainment would appear to dominate in the northern and south-western EDs. However, these are single indicators and perhaps only tell a partial story. GWPCA shows very clearly that different social structure variables dominate in different areas of Greater Dublin. In addition to analysing the dominant terms, it is useful to look at the sign of each local loading (on each local component). This identifies which variables are contrasted locally with each other in the analysis. As eigenvectors are unique only up to a multiplier of ± 1, the convention that the first loading is positive is adopted. If this is the case, 35 patterns of signs (out of a possible 256) appear in all 322 EDs, for the first component. An example map depicting 12 of the 35 different sign patterns is given in Figure 7. As example, pattern 1 (dark red) contrasts the oldest age group (negative) with the remaining seven variables (all positive); and this pattern dominates western areas of central Dublin. As example, pattern 3 (grey) contrasts social class

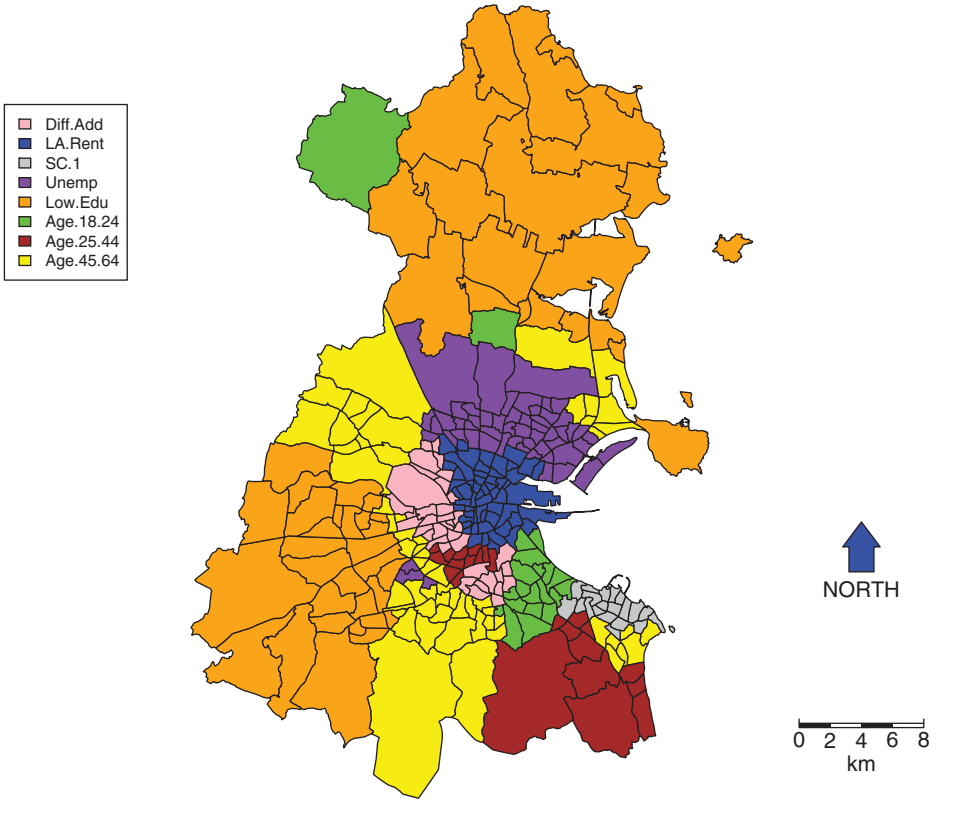

Figure 6. GWPCA ($k = 122$ and $q = 8$) output: winning variable on local component one (at each of 322 EDs).

one and the oldest age group (both negative) with the remaining six variables (all positive); and this pattern is prevalent in eastern areas of central Dublin (and some outer western areas).

Finally, it is useful to look at all eight local loadings together at each of the 322 EDs. In this respect, we use a multivariate glyph with spokes around a central hub in which the length of the spoke corresponds to the magnitude of the local loading, and its colour corresponds to the sign (in this case, red signifies negative and blue signifies positive). 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: Diff.Add is at 0 degrees (North); LA.Rent is 45 degrees (North East); SC.1 is 90 degrees (East); Unemp is 135 degrees (South East), Low.Edu is 180 degrees (South), AGE.18.24 is 225 degrees (South West), AGE.25.44 is 270 degrees (West) and AGE.45.64 is 315 degrees (North East). Figure 8 presents such a multivariate glyph map for our chosen GWPCA model (and again for the first component), where a spatial preponderance of glyphs of one colour or another, or larger spokes on the same variables provide a general indication of the structures being represented at each of the 322 ED locations. Clearly, we are presented with a difficult visualisation issue, as for the smaller EDs in central Dublin, it is hard to interpret the behaviour of numerous glyph plots. As such, a cartogram in which each zone is distorted to give each the same area (e.g. Tobler 2004) is used to represent the 322 EDs of

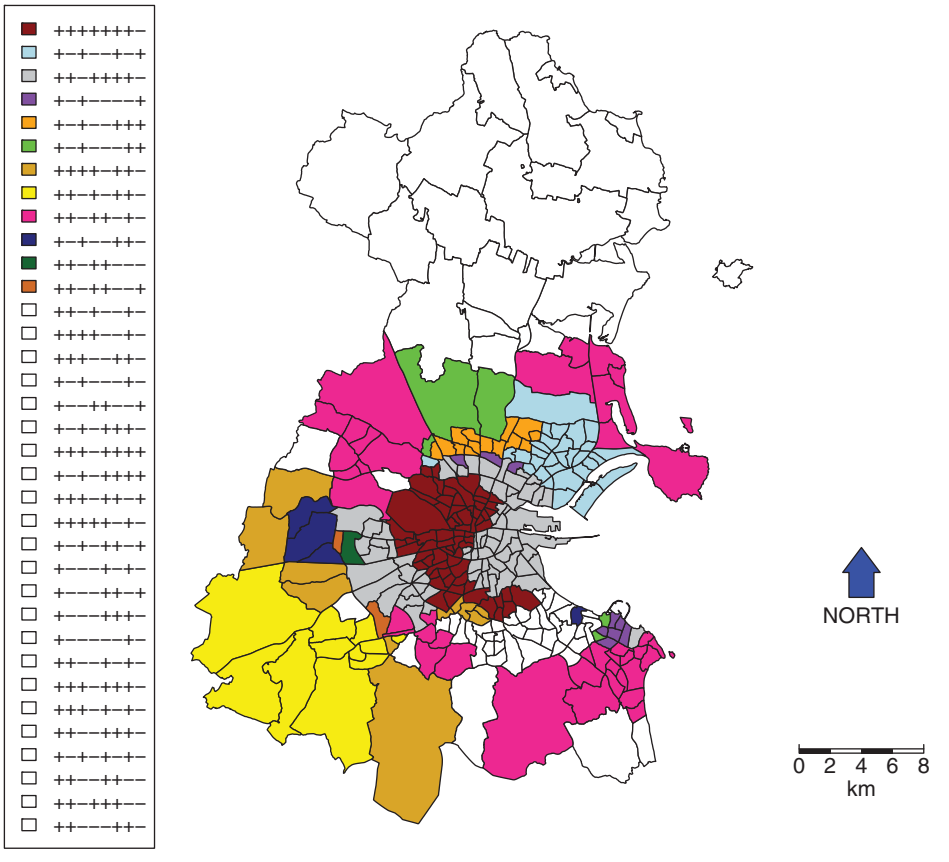

Figure 7. GWPCA ($k = 122$ and $q = 8$) output: unique combination of signs on the variables for local component one (variable order: Diff.Add, LA.Rent, SC.1, Unemp, Low.Edu, AGE.18.24, AGE.25.44 and AGE.45.64). Map only depicts 12 of the 35 different sign pattern types found; these are located at 263 of the 322 EDs.

our study area in Figure 9, where now the overall picture is much clearer. There are distinct spatial clusters of the same glyph forms. For example, many local glyphs in the south west of Greater Dublin have long blue spokes (positive loadings) in the 180 degree direction indicating a dominance of the variable measuring low education. Alternative cartograms may be useful here and as such, this is an area for further research.

4.5. Investigating collinearity in the GWR model

As PCA can be used to address collinearity in the independent data of a global regression model, similarly GWPCA should be useful in addressing collinearity in GWR, a subject that has created much debate with respect to the value of GWR as an inferential model (Wheeler 2007, 2009). We map the condition number (kappa) of our local data matrices from the same GWPCA model (Figure 10). Condition numbers above 30 are commonly used to indicate a significant collinear effect and in our local case, possible adverse effects on model inference if such data were used in a GWR fit (i.e. to explain VT, see Section 3). Clearly, many areas outside of central Dublin exhibit a high degree of collinearity and the parameters, standard errors and prediction errors from a corresponding GWR fit should be viewed with caution. We can also use GWPCA more directly by mapping the percentage of the total variance that the first (Figure 4) and last (Figure 11) components account

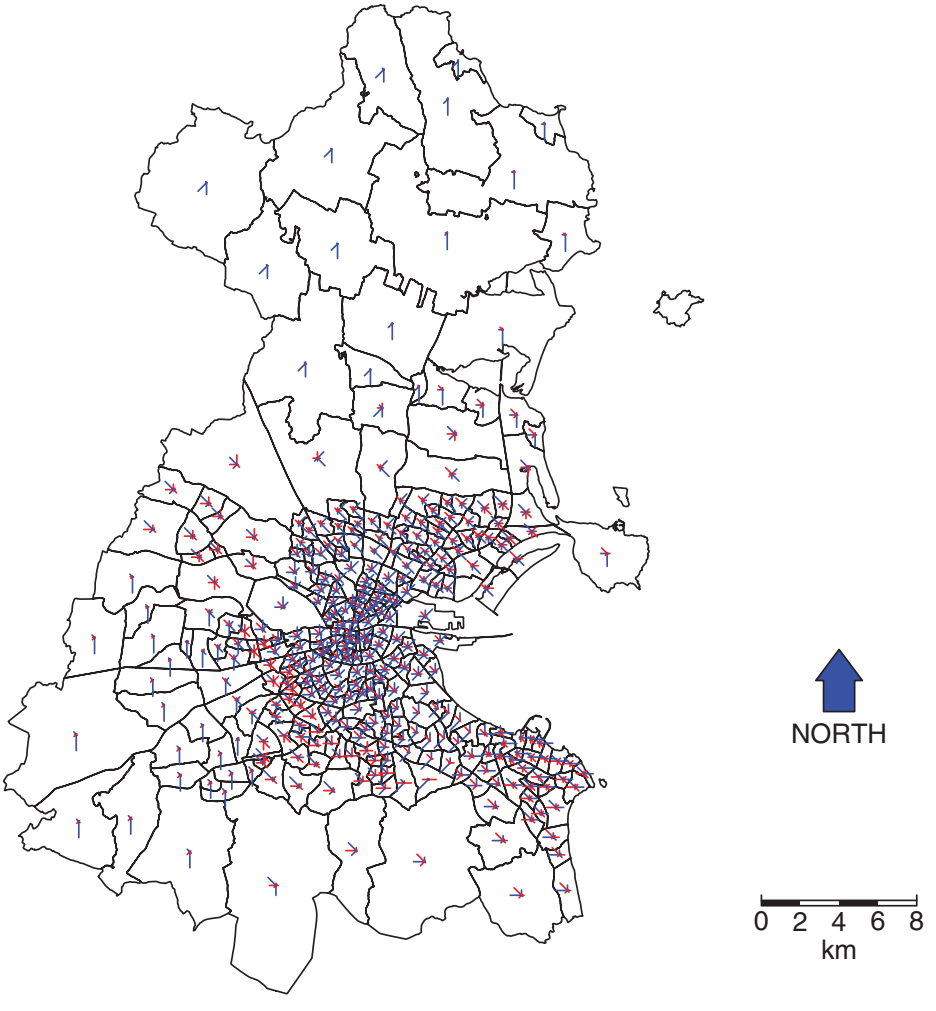

Figure 8. GWPCA ($k = 122$ and $q = 8$) output: multivariate glyphs of loadings for local component one.

for. Intuitively, the larger the first component is, the stronger the collinearity. Similarly, the smaller the last component is, the stronger the collinearity. Both effects are clearly present when such GWPCA maps are compared with the condition number map in Figure 10. This use of GWPCA appears to hold promise. Here for example, by interactively investigating the multivariate glyph cartogram maps for the first (Figure 9) and last (Figure 12) components we can locally identify the particular variables that appear to be the major causes of local collinearity.

5. Next steps

5.1. Simplification: the value of approximate forms of PCA

Principal components can be difficult to interpret at times, and the outputs of PCA may be difficult to explain to non-experts in the area. However, recent work on simplified forms of

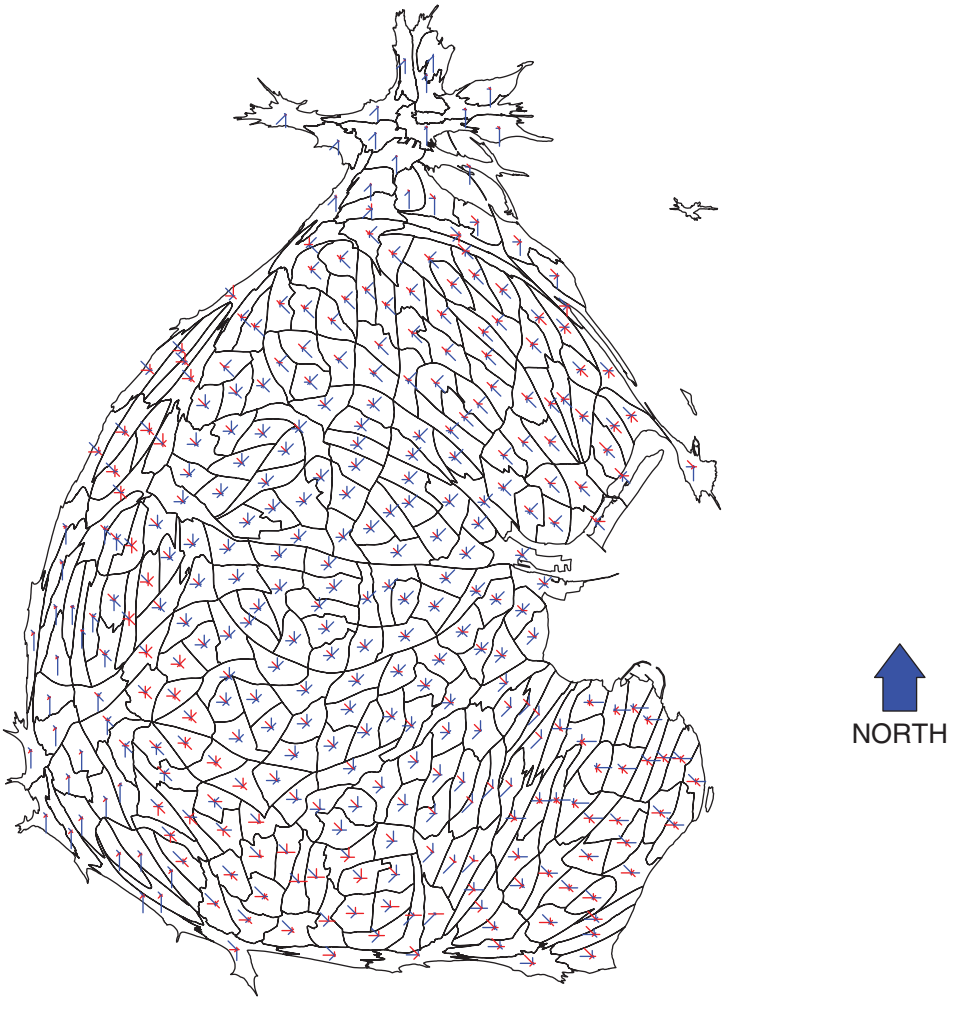

Figure 9. GWPCA ($k = 122$ and $q = 8$) output: multivariate glyphs of loadings for local component one using an equal area cartogram.

PCA has been carried out. The aim here is to constrain the values of the principal component loadings to low-valued integers. Vines (2000) presented an algorithm for this kind of analysis. GWPCA adds another dimension of complexity to the already complex task of interpreting PCA output. For this reason, attempts at simplifying PCA in a GW context will be of great use. The results of adapting Vines' approach to a GW context will be reported in a separate article.

5.2. Further extensions and uses

Just as PCA/GWPCA can be adapted or extended for a particular use, such as addressing collinearity in regression modelling, there are numerous other instances where GWPCA can replace (or complement) a standard PCA. For example, robust PCA (Rousseeuw

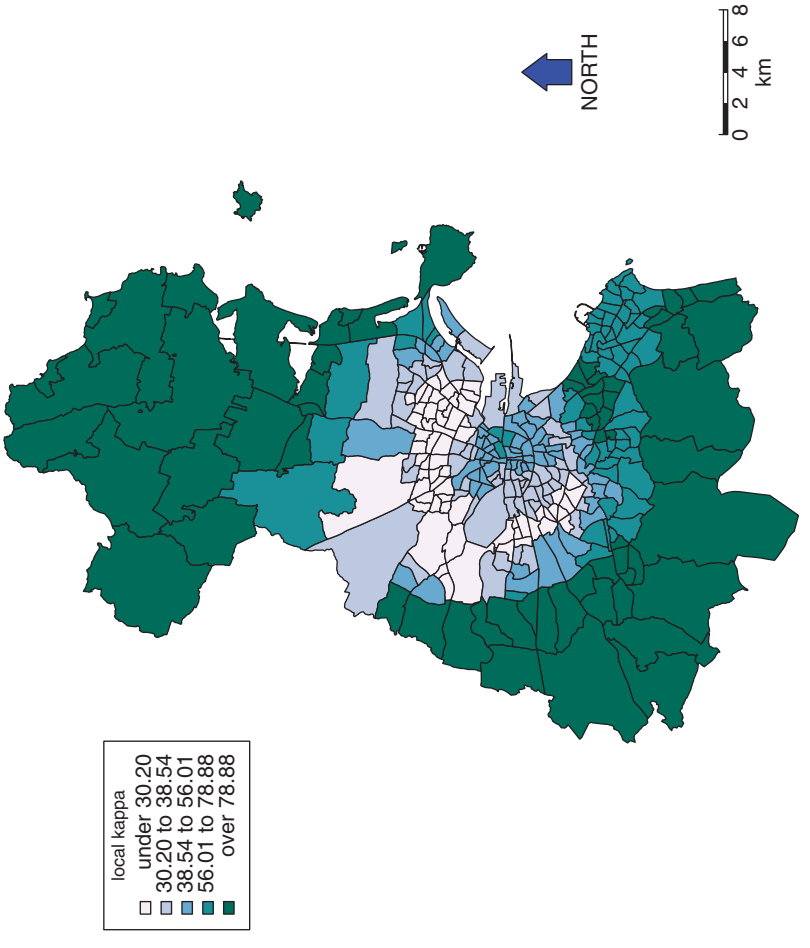

Figure 10. GWPCA ($k = 122$ and $q = 8$) output: spatial distribution of matrix condition number. Global Kappa is 16.43.

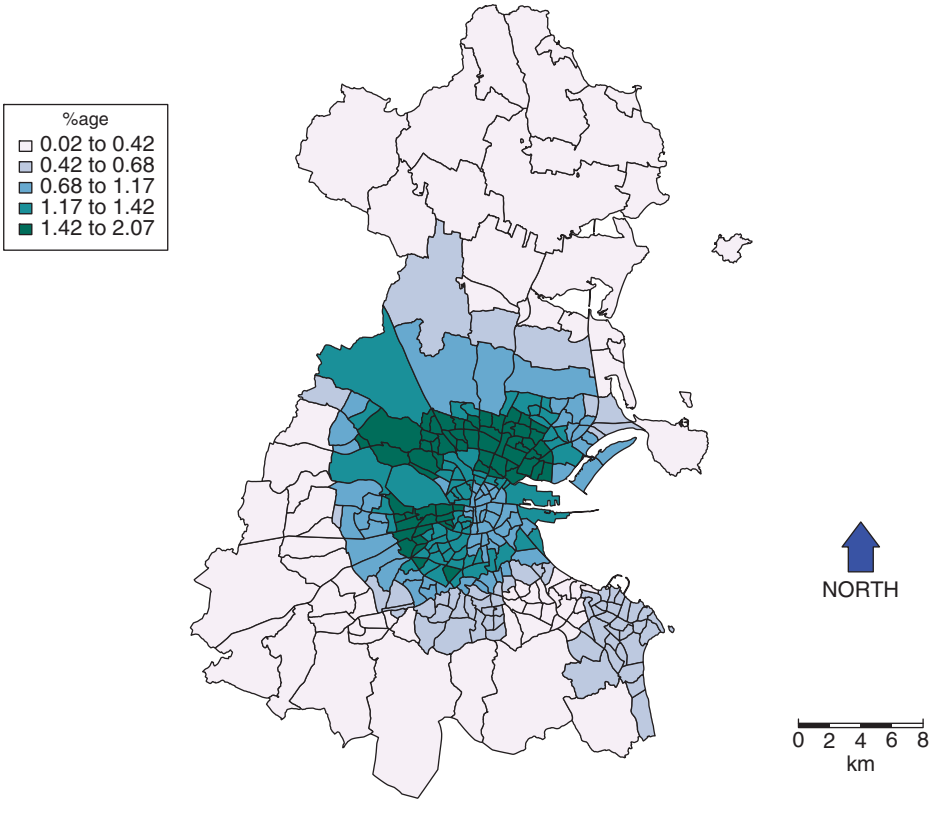

Figure 11. GWPCA ($k = 122$ and $q = 8$) output: percentage of total variation for local component eight (at each of 322 EDs). Gloabally with PCA, this percentage is 2.2%.

et al. 2006, Filzmoser *et al*. 2008) is a useful technique for detecting outliers in multivariate data sets, and a robust form of GWPCA may well help in identifying observations that are locally and spatially outlying rather than those that are globally and aspatially outlying. For example, PCA can be useful in the optimisation of sample re-design problems that need to consider both attribute- and geographic-space (Hengl *et al.* 2003) and as such, incorporating GWPCA within the same re-design algorithm may provide an improvement for a multivariate spatial process that has clear nonstationary relationship properties.

6. Conclusions

The application of a standard or global PCA presents only a partial picture in terms of variance decomposition for spatial data sets. PCA can be locally adapted to form a GWPCA technique, which whilst offering some distinct improvements presents many challenges in its specification and in the interpretation of its copious outputs. In this article, we have investigated particular issues of GWPCA calibration, testing, interpretation and visualisation, where many useful advances to the technique have been demonstrated. In addition, we have investigated a potential use of GWPCA with respect to addressing collinearity in the GWR model.

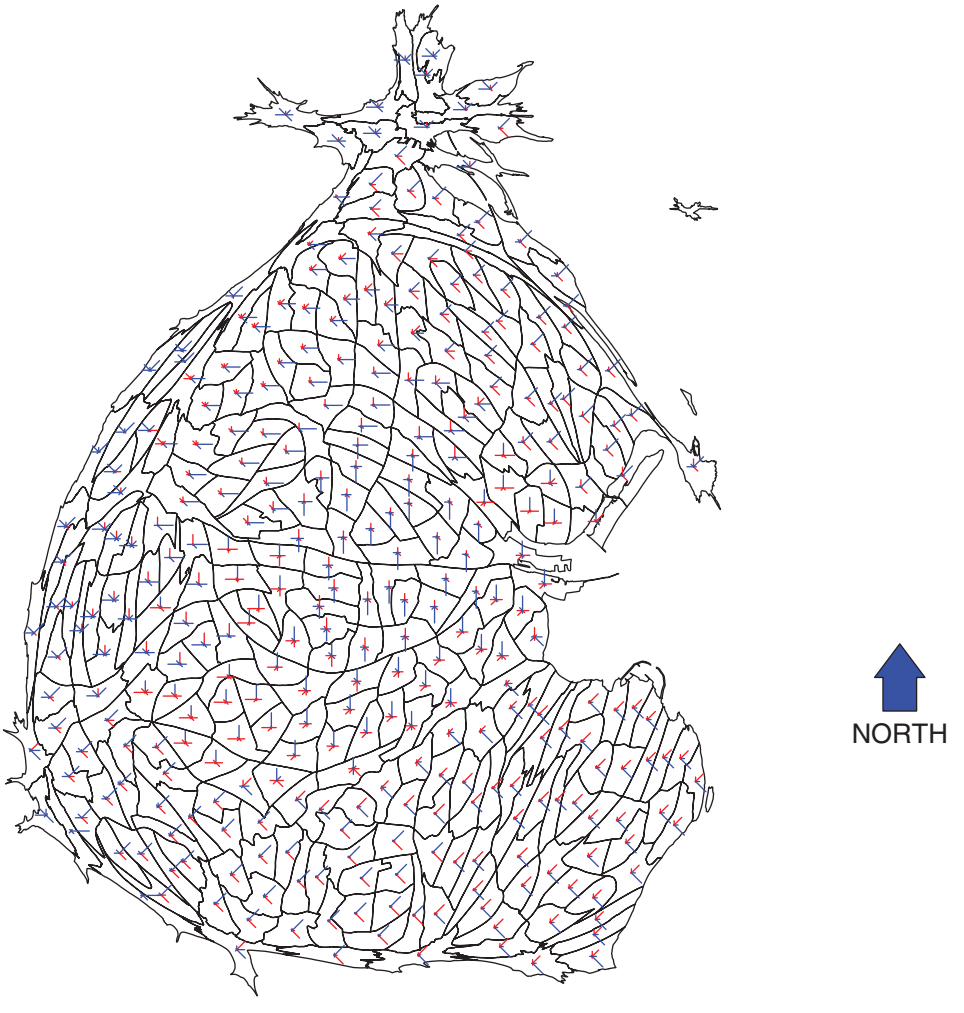

Figure 12. GWPCA ($k = 122$ and $q = 8$) output: multivariate glyphs of loadings for local component eight using an equal area cartogram.

Acknowledgements

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. The authors gratefully acknowledge this support. Professor Brunsdon's time was made available through the grant of study leave by the University of Leicester.

References

Brunsdon, C., Fotheringham, A.S., and Charlton, M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28 (4), 281–298.

Brunsdon, C., Fotheringham, A.S., and Charlton, M.E., 1998. Geographically weighted regression – modelling spatial non-stationarity. *The Statistician*, 47 (3), 431–443.

Charlton, M., Openshaw, S., and Wymer, C., 1985. Some new classifications of census enumeration districts in Britain: a poor man's ACORN. *Journal of Economic and Social Measurement*, 13, 69–96.

Chatfield, C. and Collins, A.J., 1980. *Introduction to multivariate analysis*. London: Chapman and Hall.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74 (368), 829–836.

Filzmoser, P., Maronna, R., and Werner, M., 2008. Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, 52, 1694–1711.

Fotheringham, A.S., Brunsdon, C., and Charlton, M.E., 2002. *Geographically weighted regression: the analysis of spatially varying relationships*. Chichester: Wiley.

Gould, P.R., 1967. On the geographical interpretation of eigenvalues. *Transactions of the Institute of British Geographers*, 42, 53–86.

Hägerstrand, T., 1967. *Innovation diffusion as a spatial process*. trans. A. Pred. Chicago: University of Chicago Press.

Hengl, T., Rossiter, D.G., and Stein, A., 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Australian Journal of Soil Research*, 41, 1403–1422.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 (6), 417–441 and 24 (7), 498–520.

Ihaka, R. and Gentleman, R., 1996. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299–314.

Jeffers, J.N.R., 1967. Two case studies in the application of principal component analysis. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 16 (3), 225–236.

Joliffe, I.T., 1982. A note on the use of principal components in regression. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 31 (3), 300–303.

Joliffe, I.T., 2002. *Principal components analysis*, 2nd ed. New York: Springer-Verlag.

Jombart, T., Devillard, S., Dufour, A-B., and Pontier, D., 2008. Revealing cryptic patterns in genetic variability by a new multivariate method. *Heredity*, 101, 92–103.

Kavanagh, A., 2006. Turnout or turned off? Electoral participation in Dublin in the early 21st century. *Journal of Irish Urban Studies*, 3 (2), 1–24.

Kelly, A. and Teljeur, C., 2007. *The National Deprivation Index for Health and Health Services Research*. Dublin: Department of Public Health and Primary Care, Trinity College, SAHRU Technical Report.

Lloyd, C.D., 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.

Openshaw, S., Charlton, M.E., Wymer, C., and Craft, A.W., 1987. A mark I geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, 1, 335–358.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2 (11), 559–572.

Reades, J., Calabrese, F., and Ratti, C., 2009. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environment & Planning B*, 36, 824–836.

Rousseeuw, P.J., Debruyne, M., Engelen, S., and Hubert, M., 2006. Robust and outlier detection in chemometrics. *Critical Reviews in Analytical Chemistry*, 36, 221–242.

Tobler, W., 2004. Thirty-five years of computer cartograms. *Annals of the Association of American Geographers*, 94, 58–73.

Vines, S.K., 2000. Simple principal components. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 49, 441–451.

Wheeler, D., 2007. Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environment & Planning B*, 39, 2464–2481.

Wheeler, D., 2009. Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. *Environment & Planning B*, 41, 722–742.