Principal Components Analysis: from Global to Local

Martin Charlton[1], Chris Brunsdon[2], Urška Demšar[1], Paul Harris[1], and A Stewart Fotheringham[1]

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

[2] Department of Geography, University of Leicester, Leicester, UK

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 were presented in an extensive two part paper by Hotelling (1933). In this paper we examine some problems concerning the extraction and interpretation of geographically weighted principal components (Fotheringham et al. 2002: p196-202). We initially consider the basics of principal components, the development of locally weighted principal components (LWPCA) in the exploration of local subsets in attribute space, and finally geographically weighted principal components (GWPCA). As an illustration of the use and interpretation of GWPCA we demonstrate some interpretive methods in a study of social structure in Greater Dublin. A recent application of GWPCA is provided by Lloyd's (2010) study of population characteristics in Northern Ireland and we extend some of the methodology in this paper.

PRINCIPAL COMPONENTS

Given a data matrix X with n rows representing the observations and m columns representing the variables, the variance-covariance matrix Σ is m x m with the variances in the leading diagonal and the covariances in the off diagonal elements. The trace of Σ is the total variance in the data. If the columns in X are standardised with zero mean and unit variance, the values in Σ will be those in the correlation matrix for X; its trace is equivalent to the number of columns in X.

A standard result in linear algebra states that:

$$LVL^T = R$$

where V is a diagonal matrix of eigenvalues, L is a matrix of eigenvectors and the matrix R should be a symmetric, positive definite matrix. If the matrix R is a correlation or covariance matrix denoted by Σ the eigenvalues will represent the variances of the corresponding principal components. The eigenvectors in 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. The eigenvalues can be reported as the proportion of variance accounted for by the corresponding components by dividing by tr(V).

The component scores are obtained by post-multiplying the original data values X by L; the correlation matrix for XL has 1 in each diagonal element and 0 in all the off-diagonal elements. The 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 eigendecomposition as representing a *global* model of the covariance structure of the matrix. The covariance structure of the data is assumed to be constant across the spatial extent of the study area. Thus, while the component scores may be mapped, the

eigenvectors and their associated eigenvalues are *whole-map statistics* in Openshaw et al.'s (1987) terminology.

GENERAL USES

Pearson (1901) limits himself to theoretical and some computational aspects of the fitting problem. Hotelling (1933) not only deals with the computational issues but also uses an example concerned with the reading and numerical abilities of a sample of American schoolchildren. He presents a summary which shows the eigenvalues and eigenvectors; ordering the columns in decreasing eigenvalue. He noted that the largest component 'seems to measure general ability' and the second largest 'the difference between arithmetical and verbal ability'; he also reported that these two components accounted for some 83% of the variance in the data. Little has changed in the way we approach PCA in the last three quarters of a century. However, when PCA is applied to spatial problems, there are a host of potential applications. Gould (1967) includes: measures of terrain roughness, the varying spatial nature in the connectivity of towns, orientations of physical features and transport networks, characteristics of mean information fields (Hägerstrand, 1967), classification, homogeneity of architectural features, measures of residential desirability, and the interpretation of mental maps.

Jeffers (1967) lists eight practical uses for PCA. These are:

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

Jeffers (1967) also includes examination and interpretation of the eigenvalues, interpretation of the eigenvectors, calculation of the scores, and plotting and further analysis of the scores in a list of the stages of a principal components analysis. Baxter (1974) maps principal component scores as an illustration of the possibilities offered by computer graphics.

The orthogonality of the resulting components means that they may be used in subsequent analyses. As input for principal component regression principal components provide variables which do not exhibit colinearity; variable specific parameter estimates may be obtained from the component specific estimates and the inverse of the loadings. Joliffe (1982) develops a methodology for principal component regression.

PCA is also used for data reduction: 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 1. Component scores have been used as the basis for measures of deprivation (Kelly and Teljeur, 2007) or in demographic classification systems (Charlton et al., 1985). In the first case, the analyst is interested in the values of the scores and in the second, the results of an analysis using the scores as input.

LOCALLY WEIGHTED PCA

For locally weighted principal components, the assumption of homogeneity of the correlation structure is assumed for those observations which 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' distances D instead of Euclidean distances:

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are multivariate vectors for two observations and $\mathbf{\Sigma}$ is the covariance matrix. A set of principal components is extracted for each observation in $\mathbf{X}$ using the k nearest observations in attribute space, where $k < n$. When $k = n$ the local and global component values are identical but this requires the assumption of homogeneity of the correlation structure for the whole dataset. The value of k many be supplied exogenously by the analyst or may be estimated from the data to minimise the size of the scores of those components corresponding to lower eigenvalues using either a leave-one-out or a computationally simpler holdback method.

The local eigenstructure is:

$$\mathbf{L}_i \mathbf{V}_i \mathbf{L}_i^T = \mathbf{\Sigma}_i$$

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

LWPCA provides a natural introduction to GWPCA as they are both similar in design; and it is analogous to the association between local regression in attribute space (Cleveland, 1979) and local regression in geographical space (Brunsdon et al., 1996; Fotheringham et al, 2002). Furthermore, just as LR can be used for geographical problems when the coordinate data are used as attributes, we have found that LWPCA similarly calibrated with coordinate data also has numerous applications.

GEOGRAPHICALLY WEIGHTED PCA

Following the notation of Fotheringham et al. (2002), the vector of observed variables at location i, $\mathbf{x}_i$, is assumed to have a multivariate normal distribution with mean vector $\mathbf{\mu}$ and variance covariance matrix $\mathbf{\Sigma}$.

$$\mathbf{x}_i \sim N(\mathbf{\mu}, \mathbf{\Sigma})$$

If location i has coordinates (u,v), then extending the basic model to encompass geographic effects involves regarding $\mathbf{x}_i$ as conditional on u and v, and making $\mathbf{\mu}$ and $\mathbf{\Sigma}$ functions of u and v thus:

$$\mathbf{x}_i \mid (u,v) \sim N(\mathbf{\mu}(u,v), \mathbf{\Sigma}(u,v))$$

As $\mathbf{\mu}$ and $\mathbf{\Sigma}$ are functions of u and v, this implies that each element of $\mathbf{\mu}(u,v)$ and $\mathbf{\Sigma}(u,v)$ is also a function of u and v. Thus $\mathbf{\mu}(u,v)$ and $\mathbf{\Sigma}(u,v)$ are the geographically weighted mean vector and the geographically weighted variance-covariance matrix, respectively. To obtain geographically weighted principal components, the decomposition of the geographically weighted variance-covariance matrix provides the geographically weighted eigenvalues, and the geographically weighted eigenvectors. The product of the ith row of the data matrix with the geographically

weighted eigenvectors for the *i*th location provides the *i*th row of geographically weighted component scores.

The weighting is controlled by a weighting function in an analogous manner to geographically weighted regression (GWR). The geographically weighted variance-covariance matrix is:

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

where $\mathbf{W}(u,v)$ is a diagonal matrix of weights. Typically the weights are generated from a Gaussianor bisquare function, under the control of a parameter known as the bandwidth:

$$Gaussian: \quad w_{ij} = e^{-0.5(\frac{d_{ij}}{h})^2}$$

$$Bisquare: \quad w_{ij} = (1 - (\frac{d_{ij}}{h})^2)^2$$

where *h* is the bandwidth d_{ij} is the distance between spatial locations of the *i*th and *j*th rows in the data matrix **x**. For the bisquare, w_{ij} is zero when $d_{ij} > h$. If the coordinates (u,v) are specified using a Cartesian system, then the Euclidean distance between *i* and *j* can be used in the computation; if they are in spherical coordinates, a Great Circle distance can be used.

The geographically weighted principal components for the location (u_i, v_i) can be written as

$$LVL^T \mid (u_i, v_i) = \Sigma(u_i, v_i)$$

where $\mathbf{\Sigma}(u_i, v_i)$ is the geographically weighted correlation or variance-covariance matrix for the location (u_i, v_i). These can also be referred to as a *local* PCA.

Bandwidth estimation

A major challenge in GWPCA 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 subspace. In PCA the *q*-dimensional subspace is spanned by the first *q* loadings (viewed as *m* dimensional vectors) – and is the subspace that maximised the variance of the data points projected on to that subspace. As stated earlier *q* is chosen so that this subspace contains a reasonably high proportion of the TOTAL variance. Thus components *q+1* to *m* represent the deviation from this subspace. The levels of these components therefore measure the 'goodness of fit' of the projected subplanes. Thus, a goodness of fit measure is

$$GOF_i = \sum_{j=q+1}^{j=m} s^2_{ij}$$

where GOF_i is the goodness of fit for the *ith* observation and s_{ij} is the *j*th component score for observation *i*. The total goodness of fit for the entire data set is then just

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

The *local* principal components for the *i*th location represent a similar projection, but with the corresponding loadings defined locally. The *local* linear subspace is different from the global *linear* subspace, and but we can still measure the goodness of fit in the way described above using geographically weighted principal components and scores. This statistic yields the means of finding an optimal bandwidth for GWPCA by either using a leave-one-out method or a holdback sample when computing the terms of the expression for GOF.

INTERPRETING AND VISUALISING GWPCA

GWPCA brings a set of interesting opportunities and a number of challenges. In a *global* PCA with spatial data we get, for *m* variables, *m* components, *m* eigenvalues, *m* sets of component loadings, and *m* sets of scores. It is conventional to consider 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 may be a challenge in some cases. The component scores may be mapped to yield some insights into the spatial distribution of each composite variable

In GWPCA, we get *m* components, *m* eigenvalues, *m* sets of component loadings and *m* sets of component scores for each data location in the study area. We can even obtain eigenvalues and their associated eigenvectors at locations where data has *not* been collected, although as no data exists for these locations, we cannot obtain scores.

For GWPCA we can 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 proportion of the total variance that each component accounts for, rather than the raw value. Higher values will indicate a greater variance around the component in question. It would also be useful to know whether the variance is concentrated around a few components, or whether the eigenvalues are more equally spread among the components. Computing an entropy measure might help here; if the eigenvalues are equal then the entropy measure will be $\ln(1/m)$, and if all the variance is concentrated around the first component then the entropy will be zero.

The loadings for each variable may also be mapped – this will give some indication of the variation in the spatial relationship of each variable and each component. This will show the spatial variation in the relative importance of each variable in the components, although with *m* variables there will potentially be m^2 maps to examine. However, this may be reduced by considering a subset of the components for examination in this fashion.

Fotheringham et al. (2002) suggest mapping the identification of the 'winning' variable in each local component. This is an interesting suggestion - for example in a PCA concerned with variation in social structure, GWPCA would show very clearly whether in some locations, perhaps a variable measuring single person households will dominate or whether in other locations, employment in high status occupations will dominate. A challenge arises in cases where several variables have similar high loadings, or where there are equally extreme positive and negative loadings on the component.

A valuable visual summary would be to map spatial variation using some form of multivariate glyph, perhaps with spokes around a central hub in which the length of the spoke corresponds to the magnitude of the loading, and its colour (say red or blue) corresponds to the sign. The glyphs can be

scaled relative to the spoke withthe largest absolute loading. The variables are always in the same place on the glyph, and a spatial preponderance of glyphs of one colour or other, or larger spokes on the same variables would give a general indication of the structures being represented at each location.

The biggest challenge is provided by the scores themselves. Each arises as the result of a different linear combination of the raw data. However, interpretation of the spatial variation in the scores based on a comparison with the glyph maps outlined in the previous paragraph may well yield detailed insights into the spatial variations in the local correlation structure of the data. This remains an area for further research.

To what extent can GWR recover the local spatial structures? Brunsdon et al, 2011 reports 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 eigenstructures we observe are not the result of random variation.

EXAMPLE: DUBLIN SOCIAL STRUCTURE

We use a dataset composed of a small number of variables which measure characteristics of social structure in the 322 Electoral Divisions in Greater Dublin. These have already been used in a study of voter turnout in the Irish 2004 Dáil elections (Kavanagh et al. 2006) in which the results from GWR suggest that there are distinct spatial variations in the social structure of the population of Greater Dublin. Here we use GWPCA to further uncover and interpret this spatial variation. The variables of interest are proportions of the population in each ED who are 1 year migrants, in social class 1, local authority renters, unemployed, educational attainment and in three age groups: 18-24, 25-44, and 45-64.

A global PCA reveals that three components have eigenvalues over 1, and that they account for about 69% of the variation in the data. The loadings are shown in Table 1.

Variable	PC-1 (32.5%)	PC-2 (23.8%)	PC-3 (12.8%)
1 year migrants	-0.5045	0.2853	0.1195
Local authority renters	-0.2410	-0.0989	-0.2048
Social class 1	-0.0172	0.6360	0.1598
Unemployment	-0.2229	-0.5751	-0.1311
Low educational attainment	-0.0428	-0.3926	0.5731
Aged 18_24	-0.2666	0.0311	-0.7103
Aged 25_44	-0.5321	0.1012	0.2422
Aged 45_64	0.5305	0.0908	-0.0971

Table 1: Summary of global principal components analysis

Component 1 would appear to represent older longer term residents, component 2 affluent residents, and component 3 working residents with lower educational attainment. These are 'whole map' statistics and intepretations. Whilst representing a Dublin-wide average, they may not represent local social structure particularly reliably.

The results of a GWPCA with an adaptive (bi-square) kernel with 60 observations in each local sample suggest that there is considerable variation in social structure. Figure 1 shows a map of the variation in the eigenvalue associated with the first component. The range is 4.7 to 6.5 suggesting that locally, more variation is accounted for by the first component that is the case with the global analysis. The map also suggests that the areas with the highest local eigenvalues are in the more

affluent areas of Dublin. Clearly an assessment of the local loadings would be desirable. Figure 2 shows a multivariate glyph map for the eigenvalues across all components – that the local variance is accounted for strongly by the first component is very clear across the study region.

CONCLUSION

Global PCA would appear to present only a partial picture in terms of local variation in spatial data. PCA may be local in terms of the attributes or geographical space. GWPCA would appear to offer some distinct improvements over global PCA, but presents us with some challenges for the specification of the weighting scheme and the interpretation of the results. Initial experiments have been promising.

We can take GWPCA further – PCA is a useful technique for detecting outliers in multivariate datasets, and GWPCA may well help in identifying observations which are locally rather than globally outlying. Corresponding robust forms of PCA (Daszykowski et al, 2007; Rousseeuw et al, 2006; Filzmoser at al, 2008) can be extended to robust forms of GWPCA. PCA has been shown to be useful in dealing with colinearity in the data for global regression models; similarly GWPCA should be useful in dealing with colinearity in GWR. It has also been shown that PCA is useful in the optimisation of (multivariate and spatial) sample re-design problems (Hengl et al. 2003) and as such, incorporating GWPCA within the same re-design algorithm may provide an improvement for a process that has clear heterogenic properties. Finally LWPCA may also have some uses in detecting features in LiDAR point clouds where the weighting is applied directly to the coordinate data.

BIBLIOGRAPHY

Baxter RS, 1974, On-line computer graphics in geography, International Journal of Mathematical Education in Science and Technology, 5(3), 501 — 516

Brunsdon C, Fotheringham AS, and Charlton ME, 1996, "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity", Geographical Analysis, 28(4), 281-298

Brunsdon C, Charlton M and Harris P, Simulating spatial structure in principal components, *forthcoming*

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.

Cleveland, WS, 1979, Robust Locally Weighted Regression and Smoothing Scatterplots, Journal of the American Statistical Association 74 (368): 829–836.

Daszykowski M, Kaczmarek K, Vander Heyden Y and Walczak B, 2007, Robust statistics in data analysis – a review. Basic concepts, Chemometrics and Intelligent Laboratory Systems, 85, 203-219

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

Fotheringham AS, Brunsdon C and Charlton ME, 2002, Geographically Weighted Regression: the analysis of spatially varying relationships, Chichester: Wiley

Gould PR, 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, tr. Pred A, Chicago: University of Chicago Press

Hengl T, Rossiter DG, 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

Jeffers JNR, 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 IT, 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 IT, 2002, Principal Components Analysis, 2nd edition, New York: Springer-Verlag

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, SAHRU Technical Report, Dublin: Department of Public Health and Primary Care, Trinity College

Lloyd, CD, 2010, Analysing population characteristics using geographically weighted principal components analysis: a case study of Northern Ireland in 2001, Computers, Environment and Urban Systems, doi:10.1016/j.compenvurbsys.2010.02.005

Openshaw, S, Charlton, M, Wymer, C, and Craft, AW, 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, 2, 559-572

Rousseeuw PJ, Debruyne M, Engelen S, and Hubert M, 2006, Robustness and outlier detection in chemometrics, Critical Reviews in Analytical Chemistry, 36, 221-242
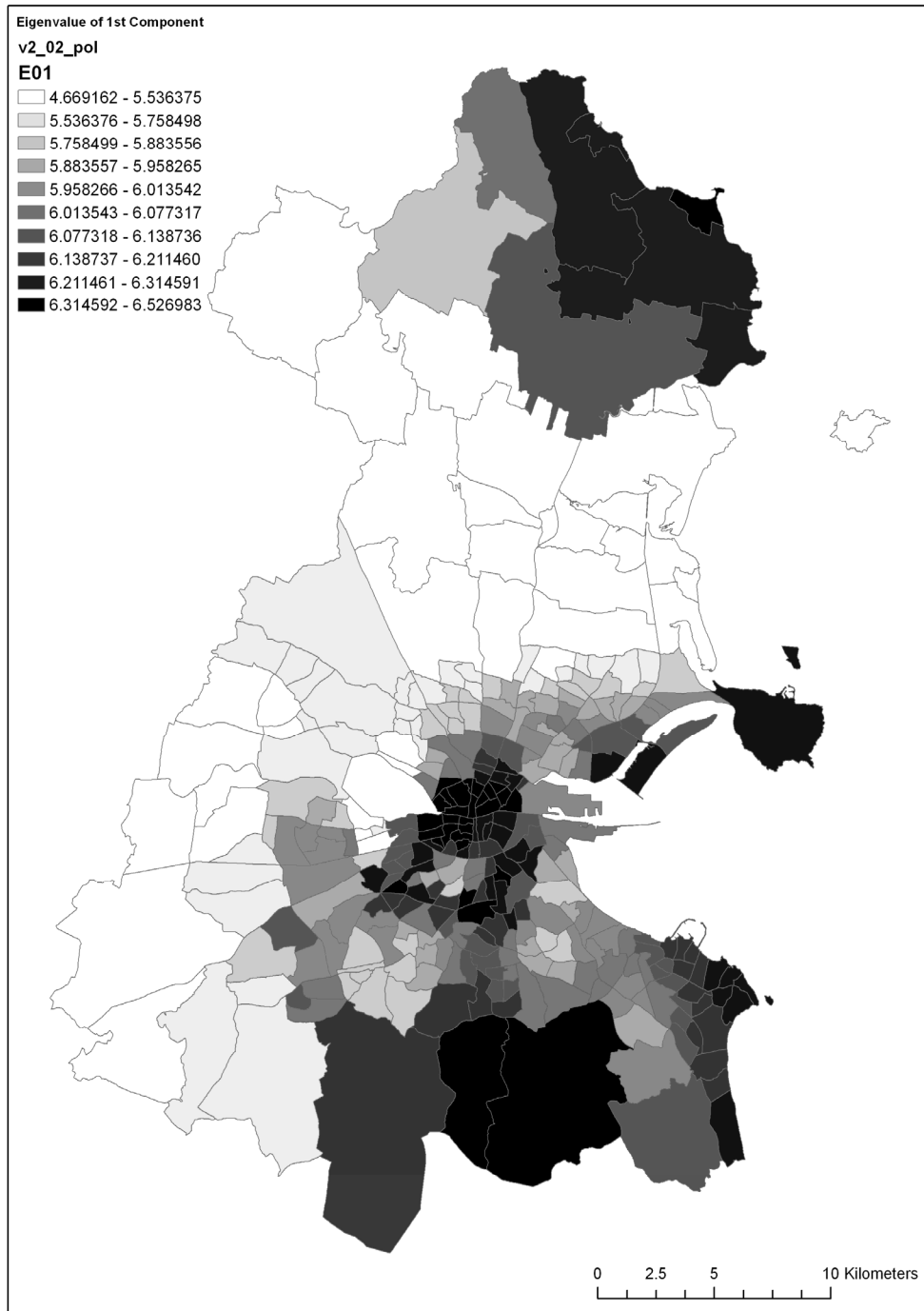
Figure 1: Map of example local eigenvalue surface in Greater Dublin
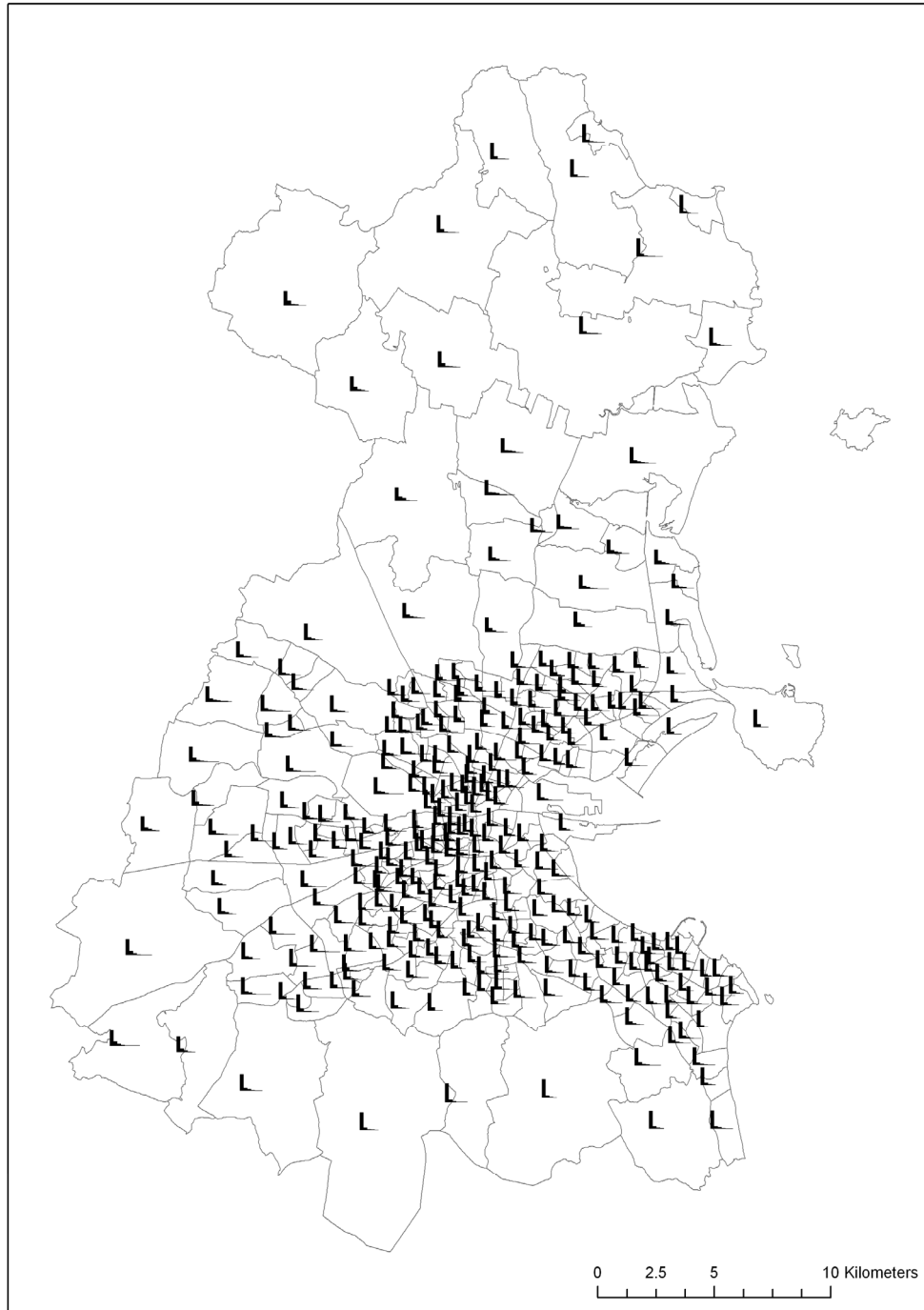
Figure 2: Map local eigenvector glyphs in Greater Dublin