Generalised additive model applied to principal component analysis of geographic data

Francisco de Asis Lopez[1],*Javier Roca-Pardinas*[2], *Celestino Ordonez.*[3].

[1]franciscoasis.lopez@uvigo.es,Department of Statistics and Operational Research. University of Vigo
[2]roca@uvigo.es, Department of Statistics and Operational Research. University of Vigo
[3]cgalan.uniovi@gmail.com, Department of Mining Exploitation and Prospecting. University of Oviedo

Abstrat

Geographically Weighted Principal Component Analysis (GWPCA) is an extension of classical PCA to deal with the spatial heterogeneity of geographical data. This heterogeneity results in a variance-covariance matrix that is not stationary but changes with the geographical location. Despite its usefulness, this method presents some unsolved issues, such as finding an appropriate bandwidth (size of the vicinity) as a function of the retained components.

In this work, we address the problem of calculating principal components for geographical data from a new perspective that overcomes this problem. Specifically we propose a scale-location model which uses generalized additive models (GAMs) to calculate means for each variable and a variance-covariance matrix that relates the variables, both depending on the spatial location. This approach does not require to calculate an optimal bandwidth as a function of the number of components retained in the analysis. Instead, the covariance matrix is estimated using smooth functions adapted to the data, so the smoothness can be different for each element of the matrix.

The proposed methodology was tested with simulated data and compared with GWPCA. The result was a better representation of the data structure in the proposed method. Finally, we show the possibilities of our method in a problem with real data regarding air pollution and socioeconomic factors.