Analyzing regional economic development patterns in a fast developing province of China through geographically weighted principal components analysis

Zaijun Li[1], Jianquan Cheng[2], Qiyan Wu[1]

1 Nanjing Normal University, Nanjing, China
2 Manchester Metropolitan University, Manchester, UK

Abstract: Understanding spatial structure of regional economic development is of importance for regional planning and provincial development strategy. Taking Jiangsu province in the economically richest Yangtze Delta as a case study, this paper aims to explore regional economic development level on a province scale. Using the census 2010, eleven variables are selected for the statistical and spatial analyses at a county level. The traditional principal component analysis (PCA) and its local version – geographically weighted PCA are employed to these analyses for the purpose of comparisons. The results have confirmed GWPCA is an effective means of analyzing regional economic development structure through mapping the local principal components. It is also concluded that the regional economic development in Jiangsu province demonstrates spatial inequality between the North and South.

Keywords: Regional economic development; spatial non-stationarity, principal component analysis; geographically weighted principal component analysis, Jiangsu.

1. Introduction

Principal Component Analysis (PCA), as a prevailing statistical analytical method, has been widely used in areas of physical science (Jeffers 1967; Harris et al. 2011) and social science (Lloyd 2010; Wu et al. 2014). The key idea underlying PCA is applying dimension reduction technique to produce few uncorrelated components from a set of original *n* correlated variables (Harris et al. 2011, 2015), while the newly created components can account for most of the variation and key trends in the original data sets. Hence, PCA has achieved an increasing popularity and naive merit in dealing with comprehensive and complex data sets collected from a variety of subject areas such as environmental and ecological sciences (e.g. Legendre and Gallagher 2001; Kaspari and Yanoviak 2009).

However the conventional or global PCA assuming constant spatial variation across the region of interests has been criticized for lacking the consideration of geographical variations and ignoring the spatial effects as the existence of spatial dependence and spatial heterogeneity is widely identified between sample units (Fotheringham et al. 2002; Kumar et al. 2012; Harris et al. 2015). Consequently, the principal components extracted from multivariate data matrix would appear to depict only a partial picture in terms of local variation in the study area (Charlton et al. 2010). As the world is not an "average" space but full of variations (Demšar et al. 2013; Harris et al. 2015), it is necessary to adapt PCA by incorporating spatial effects into the statistical analysis.

Subsequently, PCA is extended into geographically weighted PCA (GWPCA), which is local in geographic-space (Harris et al. 2010. Compared with the global PCA analysis, GWPCA is suitable to explore the impacts of geographical variation on socio-economic patterns and uncover the spatial-dynamic feature of geographical processes (Demšar et al. 2013). Thus, GWPCA is a powerful tool to reveal the changing local structure in any multivariate data sets (Lloyd 2011; Kumar et al. 2012).

In the published literature, GWPCA has been extensively applied for analyzing multivariate population characteristics (Lloyd 2010), social structure (Harris et al. 2011), soil characteristics (Kumar et al. 2012) and freshwater chemistry data (Harris et al. 2015). In these studies, GWPCA enables to reveal the spatially varying environmental and social characteristics across a study area. However, GWPCA has been rarely applied to assess spatial variability in economic systems inherently with spatially heterogeneous structure. To fill in the gap, this paper aims to explore such spatial heterogeneity present in the regional economic development structural data collected for a rapidly developing province – Jiangsu China, using the GWPCA method. The maps produced from GWPCA provide quantitative evidences and spatial details for supporting spatial plan policy and regional development strategy and help identify the spatial differentiation status of regional economic development. After this introduction, section two is focused on descriptions of the study area, data sets collected and the employed method - GWPCA. Section three is the initiative analysis of regional economic development at global level using the conventional PCA method. Then, section four is to analyze the spatial patterns of economic development at local level using the GWPCA method. The paper ends with general conclusions and preliminary discussion of spatial effects.

2. Data and Methods

2.1 The study area

Jiangsu province is located in eastern China at lower reach of Yangtze River between 30°45′ to 35°20′ N Latitude and 116°18′ to 121°57′ E Longitude (Figure 1). As a primary province of the economically richest Yangtze Delta, the province has a total area of 102,600 km^2 and a total population of 78.6934 millions and its contribution to national GDP is 10.40% in 2010 (JSB, 2011). At present, Jiangsu province administers 13 cities and 63 counties, and it is spatially divided into three parts: central (*Suzhong* in Chinese), Southern (*Sunan*) and Northern (*Subei*). In terms of regional per capita GDP in 2010, Sunan outperforms Suzhong and Subei (Figure 1).

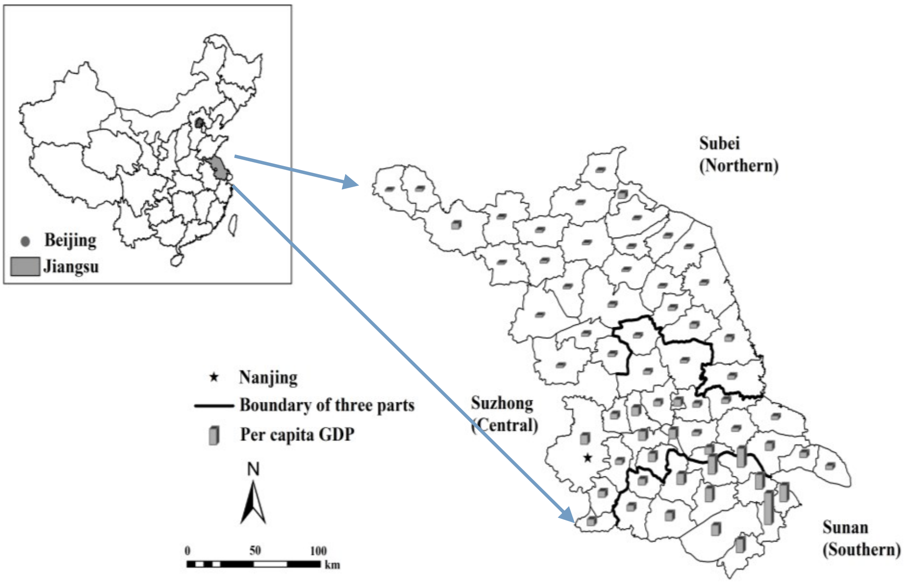

Fig.1 The study area

2.2 Data

Theoretically, regional economic development refers to economic structure, economic growth driving forces, and economic extroversion to name but a few, which can be measured by a set of statistical indices respectively (Stimson et al. 2006). For example, the driving force of economic growth can be quantified by social consumption level, fiscal revenue and land area (Aghion and Howitt 2009; Balasubramanyam et al. 2013). Economic structure is measured by the proportion of secondary industry, tertiary and fiscal expenditure to GDP, the proportion of non-agricultural workers and the amount of industrial profit tax (Li and Fang, 2014). Economic extroversion includes per capita total export-import volume, foreign investment per capita, and the ratio of foreign investment to total investment (Bassanini et al. 2001) (Table 1).

Table1. Statistical variables for measuring regional economic development level

Variable		Descriptions (unit)	
x1	Economic growth driving forces	PCSCL	Per capita social consumption level (yuan)
x2		PCFR	Per capita fiscal revenue (yuan)
x3		PCLA	Per capita land area (km^2/person)
x4	Economic structure	TRSIGDP	The proportion of secondary industry to GDP (%)
x5		TRTGDP	The proportion of tertiary to GDP (%)
x6		TRFE	The proportion of fiscal expenditure to GDP (%)
x7		TRNAW	The proportion of non-agricultural workers (%)
x8		IPTA	Industrial profit tax amount (billion yuan)
x9	Economic extroversion	PCEIV	Per capita total export-import volume (yuan/person)
x10		PCFIU	Per capita foreign investment used (yuan/person)
x11		TRFIE	The ratio of foreign investment to total investment (%)

Data source: JSB (2011)

The raw data sets for measuring the defined regional economic development pattern at county level are collected from the 2010 statistical yearbook of Jiangsu Province (JSB 2011).

2.3. Geographically weighted principal component analysis

In GWPCA, which was first coined by Fotheringham et al. (2002), the local principal components can be computed through the decomposition of local covariance. Each variable *x* has a pair of coordinates at location *i*, which is represented as X (u_i, v_i). Then, the local variance-covariance matrix is expressed as follows (equation 1):

$$\Sigma(u,v)=X^{T}W(u,v)X \qquad (1)$$

Where X is the original variables and sample unit matrix, the product of the *i*-th row of the data matrix with the local eigenvalues for the *i*-th location provides the *i*-th row of local component scores (Gollini et al. 2015); and $W(u,v)$ is a diagonal matrix of geographical weights. Further, the local principal components at location (u_i,v_i) can be expressed as follows (equation 2):

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

Where $L(u_i,v_i)$ is a matrix of local eigenvectors; $V(u_i,v_i)$ is a diagonal matrix of local eigenvalues; and $\Sigma(u_i,v_i)$ is the local covariance matrix.

In any geographically weighted method, the choice of kernel weighting function is a primary concern (Harris et al. 2015). There are diverse kernel functions provided for users to choose from such as continuous (Gaussian and exponential) and discontinuous (bi-square, tricube and box-car) functions of distance. In this paper, the bi-square kernel function is chosen due to its merits in intermediate weighting between the box-car and Gaussian functions and in producing smoothly varying results over space, which is defined as follows (equation 3):

$$w_{ij}=(1-(d_{ij}/r)^{2})^{2}\ \ if\ \ d_{ij}\leq r\ \ and\ \ w_{ij}=0\ \ otherwise \qquad (3)$$

Where d_{ij} is the geographic distance between observations i and j, r is the bandwidth and w_{ij} constitutes elements of the geographic weight matrix $W(u,v)$. The key concern is the selection of a bandwidth between a fixed distance and an adaptive distance. An adaptive bandwidth, which suits a highly irregular sample configuration (Gollini et al. 2015; Harris et al. 2015), is chosen for this study due to the nature of spatial data set used in Figure 2 (right).

Before proceeding to or interpreting the localized PCA, it is imperative to diagnose if there is any spatial non-stationarity present in the data matrix, or specifically if the geographically weighted eigenvalues from GWPCA vary significantly across space (Gollini et al. 2015; Harris et al. 2015). In statistics, this objective is usually achieved by running a Monte Carlo test (see the detailed process in Lu et al, 2014). Generally, the standard deviation (SD) of a given local eigenvalue calculated after each randomization is compared with the true SD of the same local eigenvalue. Then a significance level can be calculated from a large number of randomised distributions (e.g. 99). The results from Monte Carlo test are shown via a graph.

The GWPCA results in a series of local components variance and loading, which can be mapped to identify the spatial variation in multivariate data structure. GWPCA can assess: (*i*) how data dimensionality varies spatially and (*ii*) how the original variables influence each spatially-varying component (Gollini et al. 2015).

3. Global principal component analysis

In this case study, the selected 11 statistical variables are measured in different units, such as Yuan, Yuan/person, Km^2/person and percentage. The dissimilar magnitude between these variables may lead to biased results from PCA as the variables with the highest sample variances tend to be emphasized in the first few principal components. Hence, all the selected variables need to be standardized by subtracting its mean from that variable and dividing it by its standard deviation. Such data standardization makes each transformed variable have equal importance in the subsequent analysis.

There is another question to be answered before implementing a PCA analysis: is the sample size large enough for the statistical analysis? Is there a certain redundancy between the variables? As described before, a total number of 76 units (i.e. 13 cities and 63 counties) are observed for 11 variables. The Kaiser-Meyer-Olkin (KMO) index is run for the overall data set to detect sampling adequacy. As the KMO value is 0.717, being close to 1, the PCA can act efficiently.

The results of PCA are listed in Table 2, where the first three components with eigenvalues larger than unity totally explain up to 78.1% of variation in the regional economic development level. So, the first three components are used to explain the most variation in the data structure. Table 3 illustrates the specific components matrix with the highest absolute loadings in boldface.

The first component (PC1) accounting for 53% of variance in data dominates the structural characteristics of the regional economic development, compared with the rest components. This component (PC1) has the largest positive loading on TRNAW (0.371) and the second largest positive loading on PCFI (0.363). As such, PC1 can be used to represent main driving forces of regional economic development. The increasing foreign investment utilization provides economic growth with adequate capital sources. The growth of non-agricultural workers implies that the industrial structure is being improved as the proportion of the primary industry inclines to decreasing. Hence, these two components are related to sustainable economic growth. The variance contribution of the second component (PC2) is 14.5%, which has the largest positive loading on TRTGDP (0.513) and negative loading on TRIGDP (-0.465). As a result, PC2 can be employed to represent regional industrial structure. Comparatively, the third component (PC3) has a weak power of interpretation than the first and second as it only explains 10.6% of the variation (contrasted to 53% and 14.5% of the first and second components respectively). Accordingly, there is no further analysis of this component in detail, though it has the largest negative loading on PCLA (-0.684).

These extracted components from PCA analysis can be interpreted as new variables or indices whose statistical characteristics represent those constituent variables with the largest loadings (Jeffers 1982), while the principal components, as weighted linear combination of all variables, can be used to comprehensively assess economic development level between sample units. In Figure 2a, the higher negative PC1 scores are distributed in the centre and north, contrasting with the higher positive values relatively clustered in the Southwest. This pattern reveals that the Central and Southern areas have more influx of foreign investment and non-agricultural workers. In Figure 2b, the positive values are distributed in the Northwest, while the negative values are primarily dispersed across the Central and South. This pattern indicates that the secondary and tertiary industries relatively evenly spread across the Southern areas, but the tertiary industry accounting for most of industrial proportion are distributed in the Northern areas. All the first two principal components scores demonstrate a certain degree of geographically clustering trend across the study area.

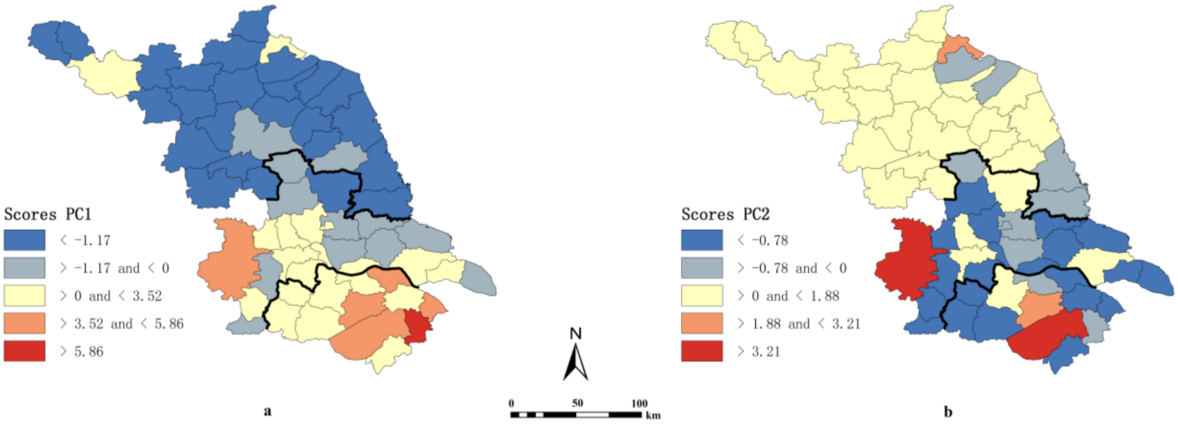

Fig. 2 Spatial distributions of PC1 and PC2

As such, PCA enables to identify the main statistical characteristics of regional economic development and reveal the intrinsic complicate interactions among the selected variables. However, all the outputs from PCA are *whole-map statistics* (Openshaw et al. 1987), which is incapable of describing local economic characteristics. In addition, the Moran's I index value for the PC1 is 0.724, which reveals a statistically positive spatial autocorrelation and as such demonstrates a highly clustering spatial pattern. Comparatively, the Moran I index value of the PC2 scores is only 0.043, demonstrating a random spatial pattern. Consequently, it is imperative to uncover the detailed local spatial variations by using GWPCA.

Table 2 Results of global PCA analysis

	Component										
	1	2	3	4	5	6	7	8	9	10	11
Eigenvalues	5.740	1.566	1.151	0.712	0.482	0.409	0.309	0.192	0.136	0.103	0.024
Standard deviation	2.396	1.251	1.073	0.844	0.694	0.640	0.555	0.439	0.369	0.322	0.155
Proportion of variance	0.530	0.145	0.106	0.066	0.045	0.038	0.029	0.018	0.013	0.010	0.002
Cumulative proportion	0.530	0.675	0.781	0.847	0.892	0.929	0.958	0.976	0.988	0.998	1.000

Table 3 The component matrix

	Component										
	1	2	3	4	5	6	7	8	9	10	11
x1	0.348	-0.199	-0.123	0.094	**0.458**	-0.147	-0.029	**0.600**	0.261	0.385	-0.076
x2	0.351	0.063	-0.106	-0.058	0.036	**0.764**	-0.071	0.009	0.182	-0.276	-0.399
x3	-0.152	-0.130	**-0.684**	**0.592**	0.100	0.048	-0.277	-0.142	-0.144	-0.079	0.076
x4	0.254	-0.465	0.250	-0.101	-0.150	-0.158	**-0.745**	-0.088	0.040	-0.180	0.056
x5	0.184	**0.513**	0.211	0.540	-0.401	-0.185	-0.122	0.247	0.274	-0.144	-0.031
x6	-0.259	0.403	-0.283	-0.408	-0.231	0.137	-0.487	0.265	-0.063	0.375	0.016
x7	**0.371**	-0.052	0.130	0.225	-0.235	0.268	0.045	-0.360	-0.127	**0.694**	0.189
x8	0.334	-0.143	-0.319	-0.128	-0.448	-0.320	0.204	0.109	-0.399	-0.015	-0.483
x9	0.296	0.364	-0.248	-0.215	0.268	-0.372	-0.090	-0.560	0.348	0.036	-0.136
x10	0.363	-0.010	-0.333	-0.235	-0.211	0.050	0.189	0.142	0.121	-0.259	**0.719**
x11	0.315	0.377	0.168	0.025	0.412	-0.002	-0.155	0.073	**-0.697**	-0.156	0.141

Note: the largest absolute loadings are shown in boldface

4. Geographically weighted principal component analysis

The GWPCA method is implemented using the GWmodel *R* package (http://cran.rstudio.com/). Firstly, a Monte Carlo test is conducted to examine whether data matrix eigenvalues are spatially varying. As shown in Figure 3, the *p*-value for testing the local eigenvalues of standard deviations from GWPCA is 0.02. This value demonstrates that the spatial invariant hypothesis of local eigenvalues is significantly rejected at the 95% level; or rather, there is a certain degree of spatial non-stationarity present in the data of regional economic development.

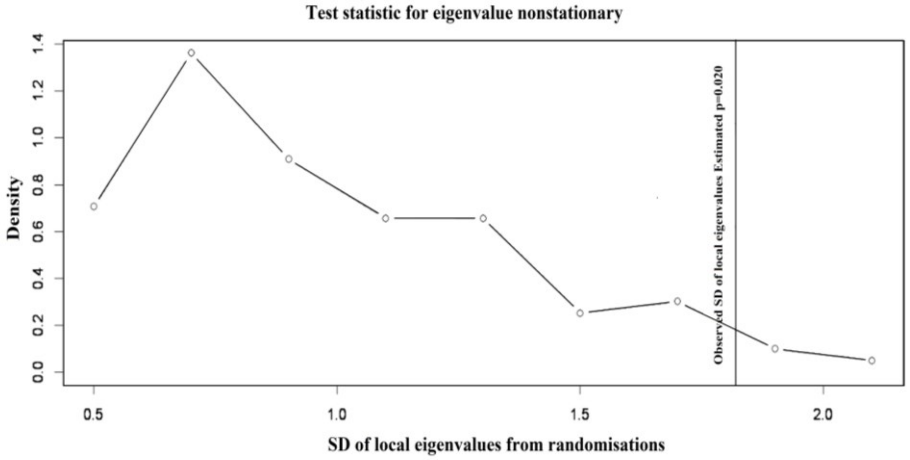

Fig.3 A Monte Carlo test of the GWPCA

Before searching for an optimal bandwidth, it is necessary to decide a prior upon the number of components to retain (Harris et al. 2015 and Gollini et al. 2015). The previous global PCA results indicate the first three components can collectively explain 78.1% of the variance in data structure. Accordingly, it is reasonable to retain three components for further GWPCA analysis. Through an adaptive bandwidth selection procedure, an optimal bandwidth of 60 km has been reached, which is chosen to run the GWPCA analysis. To be consistent with the global PCA analysis, only the first two components GWPC 1 and GWPC 2 from GWPCA will be interpreted in details for the purpose of comparisons.

As Lloyd (2010) suggested that the variables with the highest loading values and their impact intensity values can be mapped locally. Figure 4 shows the distribution of variables with the absolute highest loading from GWPC 1 (map *a*) and GWPC 2 (map *b*) respectively. On Map *a*, TRNAW dominates the most counties in the Northern areas (about 30 counties) and which is consistent with the global PCA results (Figure 2a). This pattern reveals that the newly increasing proportions of secondary and tertiary provide more employment opportunities for non-agricultural workers. As such, they become the driving force of regional economic growth for counties in Subei. PCEIV and PCFIU have the largest loading for a smaller number of areas, being 6 and 19 counties in total, mostly across the Southern and Central areas, where the regional economy has strong extroversion and their international trades are much more active than the rest. IPTA covers only 8 counties in the southwest including the capital- Nanjing city. This is because that large-scale businesses are mainly distributed in Nanjing and surrounding cities (e.g. Zhenjiang). Hence, massive tax revenue from those businesses provides lasting capital support for the economic growth of this region.

On Map *b*, GWPC 2 finds that TRSIGDP occupies 33 counties in the South and TRTGD is only active in 3 counties in the Southwest end of Suzhong. This pattern, generally being consistent with the global PCA result (Figure 2b), exhibits that the secondary industry is still the leading and pillar industry for the economic growth in this region. Comparatively, PCLA covers totally 27 counties in the North. This pattern reveals that economic development is still dependent on abundant land resources and land development. It also implies that the industry concentration degree is low in Subei and its agglomeration effect has not being achieved at current stage, although non-agricultural industry is increasing in this region. Comparatively, the second component is related to industrial structure.

Apart from the disparity in spatial distributions, these variables are also differentiated by their intensity values across the study area (Figure 4). Comparatively, Subei has higher impact intensity values in GWPC 1 and 2 than Suzhong and Sunan, and it also demonstrates continuous distributions. This differentiation can be explained by the more homogeneous economic structure in the North (Subei), where its economic development lags relatively behind the Central (Suzhong) and Southern (Sunan) areas, and the more diverse economic activities in the South, where its international trade and secondary industry play important roles. In addition, it exhibits obvious spatial spillover effects in economic growth, but which are usually confined in the inner boundary of three parts. All the analysis results reveal the underlying factors supporting economic development and the reasons why the economic development in Subei lagged behind other regions.

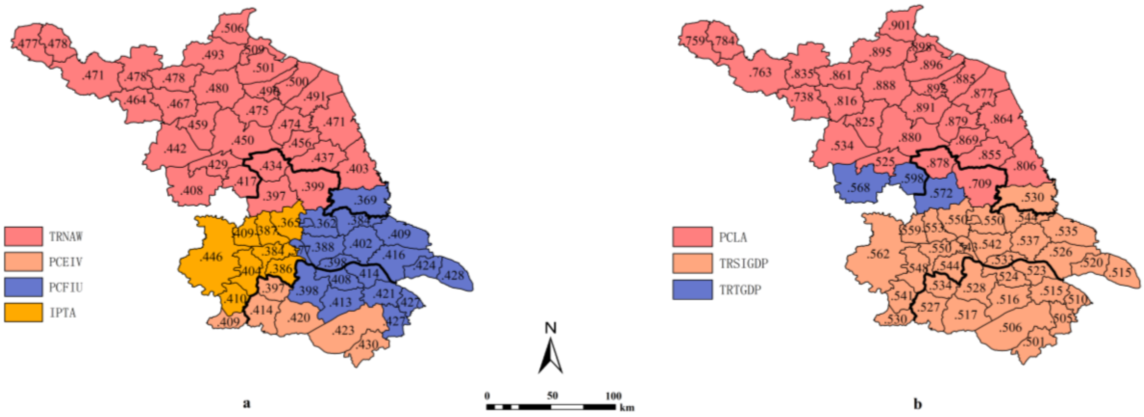

Fig.4 Variables with the largest loading and their impact intensity values: GWPC 1 (a) and GWPC 2 (b)

Compared with the outputs from global PCA, the GWPCA has exhibited its power and strength in analyzing spatial patterns of regional economic development by mapping spatial variations of each local principal component. Further, the local variance at each county explained by the calculated GWPCA 1 can be visualized by mapping as well (Figure 5), which shows a clear south-north trend with the highest percentage variances distributed in the South, intermediate level in the Central areas and the lowest values in the North. The obvious spatial clustering trend identified from the variance values in Figure 5 suggests that the interactions among these variables converge spatially. Since the cumulative percentages of variance explained by the second and the third components present similar spatial patterns as the first one, they are not interpreted again.

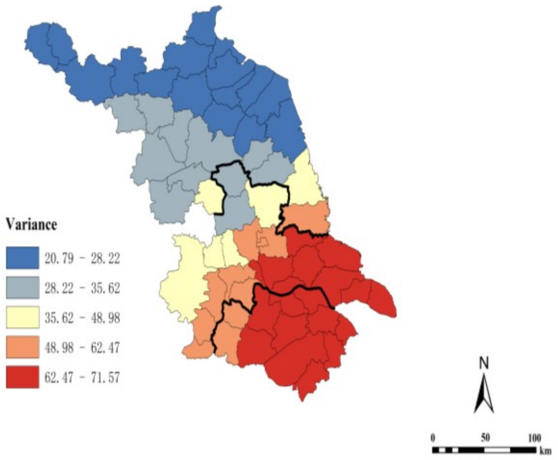

Fig.5 Percentage of variance explained by the GWPC 1

5. Conclusions

Understanding geographical variation of regional economic development is of importance for regional planning and provincial development strategy. Using the statistical data from the 2010 census of Jiangsu province at county level, this paper has comprehensively assessed the spatial variability in regional economic development using the analytical method of GWPCA. Although the global PCA is able to identify the multivariate structural characteristics, it has been criticized for ignoring spatial variations across a study area. Hence, it is natural to extend the global PCA to the variant of GWPCA. As illustrated, GWPCA produces thematic maps of local principal components, showing clear spatial structure of regional economic development. The GWPCA results confirm the hypothesis that geographical variations are present in the defined economic variables, exhibiting strong spatial differentiation between the North and South. Consequently, it can be concluded that the regional economic development structure in Jiangsu province demonstrates a strong spatial heterogeneity across its space, while this inequality can be further explored due to the spatial variations in economic development process, resource allocations, regional policies and industrial basis. Regional economic development is a complex and dynamic process. Temporal dimension should be incorporated into the GWPCA in the future, which is expected to provide more insightful findings for policy-making.

Acknowledgement:
The research is supported by National Natural Science Foundation of China (No. 41271176),Chinese Minister of Education Project of Humanities and Social Sciences (No. 12YJAZH159), and A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

Aghion, P. & Howitt, P. (2009). The Economics of Growth. The MIT Press Cambridge, Massachusetts London, England.

Balasubramanyam, V.N., Salisu, M. & David, S. (2013). Foreign Direct Investment and Growth in EP and is Countries. The Economic Journal, 106(434):92-105.

Bassanini, A. & Scarpetta, S. (2001). The Driving Forces of Economic Growth: Panel Data Evidence for the OECD Countries. OECD Economic Studies, 33:13-21.

Charlton, M., Brunsdon, C., Demšar, U., Harris, P. & Fotheringham, A.S. (2010). Principal Component Analysis: from Global to Local. The 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal, 1-10.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A.S. & McLoone, S. (2013). Principal Component Analysis on Spatial Data: An Overview. Annals of the Association of American Geographers,103(1),106-128.

Fotheringham, A.S. & Brunsdon, C. (1999). Local forms of spatial analysis. Geographical Analysis,31,340-358.

Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2002). Geographically weighted Regression: the analysis of spatially varying relationships. Chiceste:Wiley,196-202.

Gollini I, Lu B, Charlton M, Brunsdon C, Harris P (2015) GWmodel: an R Package for exploring Spatial Heterogeneity using Geographically Weighted Models. Journal of Statistical Software 63(17): 1-50.

Goodchilid, M. F. (2004). The Validity and usefulness of laws in geographic information. Annals of the Association of American Geographers,94(2):300-303.

Harris, P., Brunsdon, C. & Charlton, M. (2011). Geographically weighted principal components analysis. International Journal of Geographical Information Science, 25 (10),1717-1736.

Harris, P., Clarke, A., Juggins, S., Brunsdon, C., Charlton, M. (2015) Enhancements to a geographically weighted principal components analysis in the context of an application to an environmental data set. Geographical Analysis, 47: 146-172.

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.

Jiangsu Statistical Bureau (JSB). (2011). *Jiangsu tongji nianjing* (*Jiangsu Statistics Yearbook*). Beijing :Chinese Statistics Press.

Kumar, S., Lal, R,. & Lloyd, C.D. (2012). Assessing spatial variability in soil characteristics with geographically weighted principal component analysis. Computational Geosciences,16(3),827-835.

Li, G.D,. Fang, C.L. (2014). Analyzing the multi-mechanism of regional inequality in China. The Annals of Regional Science,52(1):155-182.

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(5),389-399.

Lu, B., Harris, P., Charlton, M., Brunsdon, C. (2014). The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models, Geo-spatial Information Science, 17(2), 85-101.

Openshaw, S., Charlton, M., Wymer, C., & Craft, A.W. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. International Journal of Geographical Information Systems, 1(4), 335-358.

Stimson, R. J., Stough, R.R, Roberts, B. H. (2006). Regional economic development. Berlin, Heidelberg: Springer.

Wu, Q.Y., Cheng, J.Q., Chen, G., Hammel, D.J. & Wu, X.H. (2014). Socio-spatial differentiation and residential segregation in the Chinese city based on the 2000 community-level census data: A case study of the inner city of Nanjing. Cities,39,109-119.

Legendre, P., Gallagher, E. (2001). Ecological meaningful transformations for ordination of species data. Oecologia, **129**(2), 271-280.

Kaspari, M., Yanoviak, S.(2009). Biogeochemistry and the structure of tropical brown food webs. Ecology. **90**(12), 3342-3351.