Identification of Heterogeneity of Social and Economic Environment of Land Uses in China

DENG Xiang-zheng[1,2*], **HUANG Wei**[1,2], **DU Ji-fu**[3], **HAN Jian-zhi**[1,2]

1. Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101; 2. Center for Chinese Agricultural Policy, Chinese Academy of Sciences, Beijing 100101; 3. College of the Humanities and Social Sciences, Graduate University of Chinese Academy of Sciences, Beijing 100049

Abstract The robust principal component analysis (RPCA) is a technique of multivariate statistics to assess the social and economic environment quality. This paper aims to explore a RPCA algorithm to analyze the spatial heterogeneity of social and economic environment of land uses (SEELU). RPCA supplies one of the most efficient methods to derive the most important components or factors affecting the regional difference of the social and economic environment. According to the spatial distributions of the levels of SEELU, the total land resources of China were divided into eight zones numbered by Ⅰ to Ⅷ which spatially referred to the eight levels of SEELU.

Key words Principal component analysis; Robust principal component analysis; Land uses; Social and economic environment; Social and economic environment of land uses

Principal component analysis (PCA) is widely used to identify the contribution of some certain factors during an integrated assessment of the regional land use environment[1]. Methodologically, PCA is capable of providing valuable information for environmental management policies benefiting the biodiversity preservation and the rational exploitation of natural and agricultural resources[2]. However, the assumptions that the observed data has a high signal-to-noise ratio, the principal components with larger variance correspond to interesting dynamics, and those with lower variance correspond to noise of PCA always limit the application of PCA, when we explored the spatial heterogeneity of social and economic environment of land uses (SEELU). By contrast, the robust principal component analysis (RPCA) rules proposed here resist outliers well and perform excellently for fulfilling various PCA-like tasks such as obtaining the first principal component vector and the first k principal component vectors as well as directly finding the subspace spanned by the first k principal component vectors. In some sense, RPCA improves the performances of the PCA algorithms significantly, when outliers are present.

SEELU is a basic element for human subsistence and connects the regional economy with social sustainable development. The evaluation for the SEELU is helpful to find out the current regional status of sustainable development and put forward the corresponding countermeasures to improve the ecological and environmental quality by carrying out an optimized land use practices. As a result, the evaluation for SEELU is popularly applied at home and abroad, and various algorithms and methodologies are used to evaluate the SEELU. There are a number of indictors used to identify the regional difference of the SEELU at a regional extent[3–4]. There are many choices for us to make, at least, those indictors from the dimensions of population growth, economic development, technical progress, infrastructure construction need to be specifically included. In addition, one more thing to be addressed here is that the inclusion or exclusion of a couple of indictors affects the final assessment results. Alfsen and S? b? identified the important basic principles behind the choice of indicators[5]. As for the integration approach, analytic hierarchy process (AHP), the common means to evaluate environment quality, is widely used in practice at present with the technical support of geographic information system (GIS). But the rest to be readdressed here is that the determinacy of the weights of factors might strongly affect the finally evaluation results of social and economic environment (SEE) at a regional extent.

As one of the most direct indicators to identify the intensity of human activities, land use constantly affects the SEELU[6]. At the same time, the form and conversion of land uses are also restricted by environment quality. So it becomes a hot topic to analyze the heterogeneity of SEELU. However, since environment is a large and multi-layer system, it is one of the biggest challenges to evaluate the SEELU using multi-level, multi-source and multi-scale data. Under the circumstances, we conducted the RPCA to solve this problem.

This paper aims to explore a reasonable method to analyze the spatial heterogeneity of SEELU by using the RPCA algorithm. The paper introduces the used data and methodology, illustrates the schemes RPCA used to derive the principal components to identify the social and economic environment conditions, and finally concludes the key findings.

Methodology

As we have addressed above, PCA supplies one of the most efficient methods to derive the most important components or factors affecting the regional differences of the SEE. As one of the multivariate statistical technique, PCA is able to analyze the dependencies existing among a set of inter-correlated variables. PCA is conducted on centered data or anomalies, and it is used to identify patterns of simultaneous variations. Its purpose is to reduce a data set containing a large

Received: August 6, 2009 Accepted: September 9, 2009
Supported by the National Scientific Foundation of China (70873118; 70821140353), the Chinese Academy of Sciences (KZCX2-YW-305-2; KZCX2-YW-326-1) and the Ministry of Science and Technology of China (2006DFB919201; 2008BAC43B01; 2008BAK47B02).

* Corresponding author. E-mail: dengxz.ccap@igsnrr.ac.cn

number of inter-correlated variables to a data set containing fewer hypothetical and uncorrelated components, which nevertheless represent a large fraction of the variability contained in the original data. These components are simply linear combinations of the original variables with coefficients given by the eigenvector. A property of the components is that each contributes to the total explained variance of the original variables. The analysis scheme requires that the component contributions occur in descending order of magnitude, such that the largest amount of variance of the first component explains the largest amount of variance of the original variables, the second explains the next largest, and so on. PCA, however, is with some limitation to be expanded to explore the spatial heterogeneity of SEELU, given that classical PCA is strongly affected by abnormal objects (outliers). In order to robustify the covariance matrix in classical PCA, the MCD-estimator and estimator of the location and shape are generally used. However, these methods might fail. In this study, a robust principal component analysis (RPCA) is investigated. RPCA is still effective, even if there are few anomalous observations.

Data and methodology

Indictor system to identify the spatial heterogeneity of SEELU The social and economic environment of land uses is a complex system. There are quite a lot of factors affecting spatial heterogeneity of SEELU at a regional extent. These factors are interactively influenced by each other. Basically, four kinds of factors at the top level, population, economy, infrastructure and technology, are included to explore the spatial heterogeneity of SEELU.

Preparation of spatial dataset and attribute dataset One of the most onerous tasks in preparing the data was to create a set of county-level observations which were consistent during the study period, since the consistency problem of county-level units generated a result of the changes of China's administrative division. As a fact, the boundaries of counties changed, and the number of counties rose over the study period. For example, China had 2 156 administrative units at the county level in 1988, whereas the number expanded to 2 733 in 2006. The organizational shifts of county-level administrative units were problematic for this study, since data within each county observational unit needed to be comparable during the study period. In order to overcome this problem, we used the geo-coding system of the National Fundamental Geographical Information System (NFGIS) [7] and a 2007 administrative map of China from the Data Center of Chinese Academy of Sciences, which included a consistent geo-coding system with that of NFGIS. Using these tools, if two counties had been subject to border shifts (e. g., one county ceded jurisdictional rights to another) ; we combined them into a single unit for the entire sample period. In case that the city core of a county had been removed from the jurisdiction of the original county-level government, we re-aggregated the municipal administrative zone back into the county-proper. In the case of large metropolitan areas (i. e., China's four provincial-level municipalities – Beijing, Tianjin, Shanghai and Chongqing, provincial capitals, and other large cities), the districts within city's administrative region were combined into a single, sample period-consistent observational unit. In this way, we ended up with a sample which includes 2 348 observational units (excluding Taiwan, Hong Kong and Macao) at the county-level that are consistent in size and jurisdictional coverage during the study period. In the rest of the paper, even though the observations included municipality district, cities and other administrative units larger and more complex than counties, for clarity, we called observations county sampling units or simply counties.

Several datasets were used to generate variables which measured the quality of SEELU of each county. Information of economy including scale, efficiency and structure for each county comes from Socio-economic Statistical Yearbook for China's Counties[8], supplemented by each province's annual statistical yearbook. The population data are from Population Statistical Yearbook for China's Counties (Ministry of Public Security of China, various years), as well as residential density, which is published by the Ministry of Public Security of China. There was a variable which measured the density of a county's infrastructure, including highway network density, road density and drainage density. Its base was a digital map of transportation and water developed by Chinese Academy of Sciences (CAS).

Schemes used to generate the map to identify the spatial clusters There were mainly eight steps by using RPCA to generate the map to identify the spatial clusters. The 1st step was to conduct singular value decomposition so as to reduce the data space to the affine subspace with dimensions. The 2rd step was to make the data points gather around the median value of the observation data. The 3rd step was to seek the first principal component with the maximal robust scale. The 4th step was to identify the data point with the data, so that the first eigenvector was mapped onto the first basis vector. The 5th step was to project the data onto the orthogonal complement of the first eigenvector. The 6th step was to repeat the 3rd step to 5th step until all required eigenvectors and eigenvalues found. The 7th step was to transform each eigenvector back to the p-dimensional space using the same reflections as in 4th step. And the final step was to link the clusters into the base map to get the final quality adjustment and thus get the clustering results of the data point to identify the spatial heterogeneity of SEELU.

Abstraction of Principal Components

Data normalization

Normalization of original data The original data (Table 1) used to calculate the index data was normalized as followings:

$$X_{ia}' = \frac{X_{ia} - \bar{X}_i}{\sqrt{S_i}} \qquad (1)$$

$$\bar{X}_i = \frac{1}{n}\sum_{a=1}^{n} X_{ia}' \qquad (2)$$

$$S_{ii} = \frac{1}{n}\sum_{a=1}^{n} (X_{ia} - \bar{X}_i)^2 \qquad (3)$$

where, $i = 1, 2, \cdots, p$ (p is indexes data) ; $a = 1, 2, \cdots, n$ (n is the number of observations).

Calculation of correlation matrix According to the following equation, the correlation matrix between variables was calculated.

$$e_{ij} = \sum_{a=1}^{n} (X_{ia} - \bar{X}_i) (X_{ia} - \bar{X}_j) \qquad (4)$$

where, $i, j = 1, 2, \cdots, p$.

The correlation matrix was then calculated as followings:

$$R = (r_{ij\,p \times p}) \qquad (5)$$

$$r_{ij} = \frac{e_{ij}}{e_{ii}e_{jj}}$$

where, e_{ij} is the deviation matrix.

Table 1 Data used for exploring the spatial heterogeneity of SEELU

Indicators	$x \pm s$
River density	8.53 ±11.90
Residential density	2.36 ±3.52
Railway density	1.09 ±2.89
Road density	7.41 ±10.56
Population number	40.38 ±78.63
Sown area of grains	99.92 ±203.00
Agricultural output value	35 468.00 ±32 170.00
Non-agricultural output value	49 416.00 ±154 017.00
Non-agricultural output value per capita	1 153.00 ±2 886.00
Agricultural output value per capita	1 156.00 ±3 184.00
Grains production per capita	583.82 ±2 234.00
Proportion of non-agricultural output value	40.44 ±21.29
Share of irrigated area to total sown area	54.11 ±36.33
Fertilizer consumption per mu	20.58 ±17.30

Abstraction of principal components It is one of the prerequisites to calculate the eigenvalues λ_i(i =1, 2, …, p) and eigenvectors l_i(i = 1, 2, …, p) according to the correlation matrix and abstract the principle components according to accumulative variance proportion. The bottom level of the proportion of the variability of the data explained by the selected principal components was around 70% -90%.

In our case study, the level of the proportion of the variability was up to 70.50%. By calculating the factor loading matrix after abstraction of principle components, we generated the factor loading matrix identifying the relationship of variance and primary factor. Factor loading is the correlation coefficient between factor and variance. Correlation matrix between factor and variance is denoted factor structure matrix. The factor structure matrix is just factor loading matrix, when factors are orthogonal. The correlation between factor load and factor variance does no longer exist after factor oblique rotation, when the common factors are not independent.

Explanation of principal components According to the methodologies and the indicator systems, a routine RPCA was conducted with the cleaned statistical data of counties of China in 2005. Based on the RPCA, five principal components were derived from the very detailed indicators. The weights of road density, residential density, drainage density and railway density on the principal component Ⅰ were the highest. The principal component Ⅰ mainly reflected the integrated situation of the four indexes, that is, the infrastructure supporting economic development, so the principal component Ⅰ was titled economic infrastructure factor. Agricultural output value per capital and grains output per capital owned the biggest weights in principal component Ⅲ. Therefore, the principal component Ⅲ mainly represented the efficiency of agricultural economic development, and it was titled efficiency factor of agricultural economic development. Proportion of irrigation area and fertilizer consumption owned the biggest weights in principal component Ⅳ. Therefore, the principal component Ⅳ mainly represented the effectiveness of agricultural technologies, and it was titled the effective factor of agricultural technologies. Total population, gross area of grains and total agricultural output value owned the biggest weights in principal component Ⅴ, so principal component Ⅴ mainly represented the scale and level of agricultural economic development, and it was titled the scale factor of agricultural economy.

Spatial heterogeneity of SEELU The above four principal components integrated the 14 variables, which identified the integrated level of the SEELU. The equation used to calculate the level of SEELU is as following:

$$\beta = \sum_{i=1}^{4} \alpha_i F_i \qquad (7)$$

where, β is the level of SEELU, α_i is weight of principal component i, and F_i is the normalized value of principal component Ⅰ by using the following equation.

$$F_1 = \frac{f_i - \min f_i}{\max f_i - \min f_i} \qquad (8)$$

where, f_i is the score of i common factor, maxf_i and minf_i are respectively the maximal and minimal values of the common factor i, and F_i is the standard value of the normalized common factor i.

$$\alpha_i = \lambda_i / \sum_{i=1}^{p} \lambda_i \qquad (9)$$

where, α_i is the weight of common factor Ⅰ, and λ_i is the eigenvalue of common factor i.

Eight levels of the SEELU were identified by the calculation. According to the spatial distributions of the levels of SEELU, the total land resources of China were divided into eight zones numbered by Ⅰ to Ⅷ and spatially referenced to the eight levels of SEELU. Zone Ⅰ with an area of 1.00% of the total land resources was mainly distributed in coastal regions or around the mega cities, e.g., Liaoning, Shandong, Jiangsu, Guangdong, Beijing, Shanghai, Tianjin, Chengdu, Shenyang, Wuhan, etc. Zone Ⅱ, Ⅲ and Ⅳ with an area of 11.37% of the total land resources were mainly distributed in eastern coastal areas including Liaoning peninsula, Shandong peninsula, Huabei plains, middle and lower reaches Plains of Yangtse River, Yangtse River Delta, Pearl River Delta, Sichuan Basin and Guanzhong Basin, which are developed regions with densely population distribution and good infrastructure. Zone Ⅴ and Ⅵ with an area of 27.10% of the total land resources were mainly distributed in eastern regions covered by hills and low mountains, and these regions were featured by geophysical conditions redistricting the economic development to some extent. Zone Ⅶ and Ⅷ, an arid and sub-arid area occupying 47.46% of the total land resources of China, were mainly distributed in the 1st and 2nd grades of topography of China, and these regions were featured by physical conditions redistricting the economic and social development at the regional extent. The forestry and animal husbandry took the main parts in the regional industrial structure.

Conclusion and Discussion

It is of significance to identify the spatial heterogeneity of social and economic environment of land uses for exploration of the scientific and practical land use plans at regional extent. A lot of indicators, from the domains of demography, economy, technology and infrastructure, were identified to evaluate the regional difference of the SEELU in China. As a basic indicator to identify the social and economic environment of land uses, SEELU is characterized with an obvious spatial heterogeneity. In our study, five principal components were derived from the very detailed indicators. Eight grades of the social and economic environment of land uses were identified by the integrated assessment. In this sense, the RPCA-based assessment for the social and economic environment of land uses is of importance within the context of a clear hierarchy of planning policy for land uses, and it is generally consistent with and complements national policy and region-wide policy.

References

[1] MORIN G, FORTIN JP, SOCHANSKA W, et al. Use of principal component analysis to identify homogeneous precipitation stations for optimal interpolation[J] . Water Resour Res, 1978, 15(6) : 1841 –1850.

[2] LASAPONARA R. On the use of principal component analysis (PCA) for evaluating interannual vegetation anomalies from SPOT/VEGETATION NDVI temporal series[J] . Ecological Modelling, 2006, 194(4) : 429 –434.

[3] DENG XZ, LIU JY, ZHUANG DF, et al. A typical method based on remote sensing and GIS for integrated environmental assessment and its application in China[C] . Kanazawa, Japan: Proceedings of EMEA01 in Kanazawa, 2001.

[4] GAO ZQ, DENG XZ. Analysis on spatial features of LUCC based on dataset of land use and land cover change in China[J] . Chinese Geographical Science, 2002, 12(2) : 107 –113.

[5] ALFSEN KH, HANS VS. Environment quality indicators: background, principles and examples from Norway[J] . Environmental and Resource Economics, 1993(3) : 415 –435.

[6] WOLMAN MG. Population, land use, and environment: a long history[M] . Washington, DC: National Academy Press, 1993: 15 –29.

[7] National Fundamental Geographic Information System. Geocoding system of administrative zones of in 1995 (GB 2260 –1995) [M] . Beijing: National Fundamental Geographic Information System, 2000.

[8] National Bureau of Statistics of China(中华人民共和国国家统计局) . China statistical yearbook 2006(中国统计年鉴 2006) [M] . Beijing: China Statistics Press(北京: 中国统计出版社) , 2006.

[9] DING H(丁辉) , HUANG L(黄磊) , XIE K(谢柯) , et al. Evaluation of land ecological security in Sichuan Province(四川省土地生态安全评价) [J] . Journal of Anhui Agricultural Sciences (安徽农业科学) ,2009, 37(33) :303 –304, 339.

[10] DENG XZ(邓祥征) , ZHAN JY(战金艳) , SU HB(苏红波) . Simulation and analysis of land system structure changes in Huang –Huai –Hai plain area(黄淮海平原土地系统结构变化的模拟与分析) [J] . Agricultural Science & Technology(农业科学与技术) , 2007, 8(3 –4) :45 –52.

[11] DUAN QW(段清伟) , ZHANG RQ(张润清) , CAO WW(曹文文) . (河北省农村土地承包经营权流转的影响分析) [J] . Animal Husbandry and Feed Science (畜牧与饲料科学) , 2009, 30(9) : 174 –175.

[12] MENG M(孟敏) , LI D(李丁) . Analysis on temporal variation tendency of cultivated land area in Tianshui City Gansu Province(甘肃省天水市耕地面积时序变化分析) [J] . Journal of Anhui Agricultural Sciences(安徽农业科学) ,2009, 37(34) :214 –216, 265.

Responsible editor: ZHANG Ming-ming Responsible translator: ZHANG Ming-ming Responsible proofreader: WU Xiao-yan

中国土地利用社会经济环境综合评价方法

邓祥征[1,2*] , 黄 维[1,2], 杜继福[3], 韩健智[1,2] (1. 中国科学院地理科学与资源研究所, 北京 100101; 2. 中国科学院农业政策研究中心, 北京 100101; 3. 中国科学院研究生院人文学院, 北京 100049)

摘要 稳健主成分分析(RPCA) 法可用于评价土地利用的社会经济环境。该文展示了 RPCA 法在土地利用社会经济环境综合评价中的应用。利用 RPCA 法并基于土地利用的社会经济环境的区域分异特征将中国土地利用社会经济环境划分为 8 个分区, 分别代表土地利用社会经济环境的 8 个级别。研究表明, 采用 RPCA 法不易受到土地利用社会经济环境各成分要素异常值影响, 所获得的土地利用社会经济环境分区较为科学与合理。由此可见, 应用 RPCA 法可以有效地提取土地利用社会经济环境表征指标, 可应用于提炼土地利用社会经济环境区域分异特征。

关键词 主成分分析; 稳健主成分分析; 土地利用; 社会经济环境; 土地利用社会经济环境

基金项目 国家自然科学基金(70873118; 70821140353) ; 中国科学院知识创新工程项目 (KZCX2-YW-305-2; KZCX2-YW-326-1) ; 科技部国际合作项目 (2006 DFB919201) 和国家科技支撑项目 (2008 BAC43 B01; 2008BAK47B02) 。

作者简介 邓祥征(1971 –) , 男, 山东日照人, 博士, 副研究员, 从事土地系统变化与效应及区域环境变化研究。* 通讯作者。

收稿日期 2009-08-06 修回日期 2009-09-09

(上接第 137 页)

点调查 20 株, 分点调查记载病害名称、有病害株数、病死株数; 虫害名称、有虫株数、虫数。同时, 记载病害和虫害危害症状。从川芎病害田块中采集病株, 用病株的病健部按常规方法分离病原菌, 并进行纯化培养, 鉴别病原菌种类, 并观察其培养性状。采集川芎害虫样品, 在室内鉴定害虫种类。

[结果] 川芎的主要病害有根腐病、白粉病、斑枯病。在苓种阶段, 以根腐病、白粉病危害最重, 大田期以根腐病危害严重。根腐病 4 月中下旬至6 月中旬进入盛发期, 在苓种田块 7 月中旬达到发病高峰, 连作、排水不良、氮肥施用量大的田块发病严重; 白粉病一般在 5 月下旬开始发生, 6 ~7 月高温高湿时发病严重; 斑枯病在 5 月上旬开始发生, 5 月底 6 月初川芎即将收获时大田发生普遍, 7 ~8 月苓种叶片老熟后发病严重。川芎的主要虫害有茎节蛾、斜纹夜蛾、蛴螬、红蜘蛛和种蝇等。在苓种阶段, 主要有茎节蛾、斜纹夜蛾、蛴螬和种蝇, 在大田期以种蝇危害最为严重。其中, 茎节蛾以 6 月中旬到 7 月中旬的二、三代发生危害最为严重; 斜纹夜蛾在 7 ~8 月大发生; 蛴螬7 月中旬进入为害盛期; 叶螨在高温低湿的 6 ~8 月发生危害严重; 种蝇在川芎整个生育期均可造成危害, 在春、秋季发生最为严重。同时在水肥充足条件下发生普遍, 尤其是粪肥施在表面时发生严重。

[结论] 明确了川芎主要病虫害的发生危害规律, 为川芎重要病虫害的防治提供了科学依据。

关键词 川芎; 病害; 虫害; 发生规律

基金项目 “十| 五”国家科技支撑计划(2006BA109B04-05) ; 四川省青年基金(09ZQ026-039) ; 四川省农业科学院优秀论文基金项目 。

作者简介 曾华兰(1969 –) , 女, 重庆万洲人, 硕士, 副研究员, 从事中药材等经济作物病虫害防治及评价研究。* 通讯作者。

收稿日期 2009-12-16 修回日期 2010-01-11