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

DENG Xiang-zheng1,2, HUANG Wei1,2, DU Jie-fu3, HAN Jian-zhi1,2
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 level of SEELU, the total land resources of China were divided into eight zones numbered by 1 to 8, which spatially referred to the eight levels of SEELU.

Keywords 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. 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. 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 performance of the PCA algorithm 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 optimal 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 indices used to identify the regional difference of the SEELU at a regional extent. There are many choices for us to make, at least, those factors from the dimensions of population growth, economic development, technological 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 indicators affects the final assessment results. Aasen and Stadler identified the important basic principles behind the choice of indicators. As for the integration approach, analytic hierarchy process (AHP), the common means to evaluate environment quality, widely used in practice, present with the technical support of geographic information system (GIS). But the rest to be addressed here is that the determinant of the weights of factors might strongly affect the final 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 SEE. At the same time, the form and conversion of land uses also are restricted by environment quality. So it becomes a hot topic to analyze the heterogeneity of SEELU. However, since environment is a large and multilayer system, it is one of the biggest challenges to evaluate the SEELU using multilevel, multisource and multiscale 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 interrelated 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
number of intercorrelated variables in 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 var-
iables, the second explains the next largest, and so on. PCA,
however, is with some limitation to be expanded to explain
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 gener-
ally used. However, these methods might fail. In this study,
a robust principal component analysis (RPCA) is investigat-
ed. RPCA is still effective, even if there are anomalous ob-
servations.

Data and methodology

Indicator system to identify the spatial heterogeneity of
SEELU[3]. 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, in-
frastructure and technology, are included to explain the spa-
tial heterogeneity of SEELU.

Preparation of spatial dataset and attribute dataset[4]. One
of the most essential tasks in preparing the data was to create
a set of county-level observations which were consistent dur-
ing the study period, since the consistency problem of county-
level units generated a result of the changes of China’s admin-
istrative division. As a fact, the boundaries of counties
changed, and the number of counties rose over the study peri-
d. For example, China had 2156 administrative units at the county level in 1988, whereas the number expanded to 2733
in 2006. The organizational shifts of county-level administra-
tive units were problematic for this study, since data within
each county observational unit needed to be comparable dur-
ing the study period. In order to overcome this problem, we
used the geo-coding system of the National Fundamental Geo-
ographical Information System (NGFIS)[5] and a 2007 admin-
istrative map of China from the Data Center of Chinese Acad-
emy of Sciences, which included a consistent geo-coding sys-
tem with that of NGFIS. Using these tools, if two counties had
been subject to border shifts (e.g., one county ceded jurisdic-
tional rights to another), we combined them into a single u-
nit for the entire sample period. In case that the city core of
a county had been removed from the jurisdiction of the orig-
inal 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, sam-
ple period-consistent observational unit. In this way, we en-
ded up with a sample which includes 2348 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 districts, 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[6], 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 den-
sity, 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. As 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[7]. 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
2nd step was to make the data points gather around the medi-
an 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 points 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 comple-
ment of the first eigenvector. The 6th step was to repeat the
3rd step to 5th step until all required eigenvectors and eigen-
values 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 het-
eregogeneity of SEELU.

Abstraction of Principal Components

Data normalization

Normalization of original data[8]. The original data (Table 1)
used to calculate the index data was normalized as followings:

\[ X_i = \frac{X_i - \bar{X}}{\sigma} \]  
\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \]  
\[ \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2} \]

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

Calculation of correlation matrix[9]. According to the fol-
lowing equation, the correlation matrix between variables was calcu-
lated

\[ c_{ij} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) (X_j - \bar{X})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{j=1}^{n} (X_j - \bar{X})^2}} \]

where, \( i, j = 1, 2, \ldots, p \).

The correlation matrix was then calculated as followings:

\[ R = \left( r_{ij} \right) \]
Table 1 Data used for exploring the spatial heterogeneity of SEELU

<table>
<thead>
<tr>
<th>Indicators</th>
<th>x 2 x 2s</th>
</tr>
</thead>
<tbody>
<tr>
<td>River density</td>
<td>8.53</td>
</tr>
<tr>
<td>Residential density</td>
<td>2.36</td>
</tr>
<tr>
<td>Railway density</td>
<td>1.09</td>
</tr>
<tr>
<td>Road density</td>
<td>7.41</td>
</tr>
<tr>
<td>Population</td>
<td>40.38</td>
</tr>
<tr>
<td>Sown area of grains</td>
<td>99.92</td>
</tr>
<tr>
<td>Agricultural output value</td>
<td>35.48</td>
</tr>
<tr>
<td>Non-agricultural output value</td>
<td>49.41</td>
</tr>
<tr>
<td>Non-agricultural output value per cap</td>
<td>1.53</td>
</tr>
<tr>
<td>Agricultural output value per cap</td>
<td>1.56</td>
</tr>
<tr>
<td>Grains production per cap</td>
<td>583.82</td>
</tr>
<tr>
<td>Proportion of non-agricultural output value</td>
<td>40.44</td>
</tr>
<tr>
<td>Share of irrigated area to total sown area</td>
<td>54.11</td>
</tr>
<tr>
<td>Fertilizer consumption per mu</td>
<td>20.58</td>
</tr>
</tbody>
</table>

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:

\[ \beta = \sum \alpha_i F_i \]

where, \( \beta \) is the level of SEELU, \( \alpha_i \) is weight of principal component \( i \) and \( F_i \) is the normalized value of principal component by using the following equation:

\[ F_i = \frac{t - \text{min}}{\text{max} - \text{min}} \]

where, \( t \) is the score of common factor, \( \text{max} \) and \( \text{min} \) are respectively the maximal and minimal values of the common factors, and \( F_i \) is the standard value of the normalized common factors.

\[ \alpha_i = \frac{\lambda_i}{\sum \lambda_i} \]

where, \( \alpha_i \) is the weight of common factor \( i \), and \( \lambda_i \) 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 areas 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, midland and lower reaches plains of Yangtze River, Yangtze River Delta, Pearl River Delta, Sichuan Basin and Guanzhong Basin, which are developed regions with densely populated 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 geo-physical conditions redirecting 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 redirecting 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 exploitation 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 PCA-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


