Spatial Clustering and Outlier Analysis for the Regionalization of Maize Cultivation in China

Hu WANG, Xiaodong ZHANG, Shaoming Li, Xiaomei SONG
Department of Geographic Information Science
China Agricultural University
No. 17 Tsinghua East Road, Beijing 100083
P.R. CHINA
wanghugigi@gmail.com
Corresponding Author: Xiaodong ZHANG
zhangxd@cau.edu.cn

Abstract: Regionalization has been the foundation of large-scale plantation and local optimization for crop cultivation. Current regionalization approaches practiced mainly rely on qualitative analysis and heuristic methods, which cannot meet the increasingly challenging demands. In this paper, we demonstrate the use of spatial clustering method on the regionalization of crop cultivation, with the maize growing in China as an example. In the proposed method, we adopt four indicators [that is, elevation, effective accumulated temperature (EAT), precipitation and yield] which are the major factors reflecting the maize cultivation differences. In addition, by taking into account the spatial information of counties in the clustering process, we achieve a more spatially coherent clustering result. As a post-processing step, adjustment with the help of a Geographic Information System (GIS) eliminates regions that appear inconsistent with the vicinity. With the proposed approach, we classify the 2,831 counties of China into 7 regions, and show that the result is highly consistent with the conventional regionalization of maize cultivation in China. This result proves the feasibility of our approach, and suggests its possible application on other crops. Furthermore, we carry out outlier analysis for each of the regions to identify the counties that show abnormal behaviors in maize cultivation, and further analyze the possible causes. This study provides valuable information for cultivation region selection in large-scale crop plantation.

Key-Words: Spatial clustering, regionalization, maize-growing region, outlier analysis

1 Introduction

Regionalization of crop cultivation, which provides the basis for optimal resource utilization and production planning in agriculture, is one important study in geography. As a part of agriculture planning, regionalization focuses on ways of partitioning plantation area according to the adaptation of crops to local conditions, with the aim of providing scientific basis for large-scale plantation and local optimization [1]. However, the regionalization approaches adopted in practice today in China still heavily rely on heuristics and qualitative arguments, which lack accuracy and cannot meet the requirement of providing guidance for precise crop production [2-3]. Little attention has been paid to the study of its theory and practice from either the geography and agriculture community in China.

In this paper, we demonstrate the use of spatial clustering methods on the regionalization of crop cultivation. Clustering is the assignment of objects into groups (called clusters) so that objects from the same cluster are more similar to each other than objects from different clusters [4]. Clustering analysis is derived from a number of diverse fields, including data mining, statistics, biology and machine learning. As an unsupervised learning method, clustering does not rely on the predefined classes and class-labeled training examples. Cluster analysis has been widely used in numerous applications, including data analysis, pattern recognition, image processing and market research[5-7].

Spatial cluster analysis is an extension of clustering methods to spatial data, aiming to discover spatial distribution patterns among data attributes. It has been widely applied in applications such as pollution monitoring, climate change monitoring and disastrous weather forecasting. The subjects of

spatial clustering range from raster data to vector data (including points, lines, faces and discrete or continuous polygons) [8]. However, most current approaches of spatial clustering concern with points and discrete polygons [9-10]; while continuous polygon, as an important data type in geography, is not placed enough emphasis on. The clustering of continuous polygons often requires the entities within the same clusters to be similar in attributes, while contiguous in geographic locations. This is consistent with the requirement of regionalization.

So far as we know, studies on the application of clustering analysis in regionalization of crop cultivation [11-14] have been conducted, but they mainly focus on attributes clustering. In this paper, we demonstrate a novel spatial clustering method of continuous polygons for the application of regionalization of maize cultivation. In addition, we apply outlier analysis to identify regions that show abnormal characters and show their possible causes.

The paper is organized as follows: Section 1 introduces the selection and preprocessing of indicators used in clustering. Section 2 presents the proposed spatial clustering algorithm and its results when applied to maize cultivation regions. Section 3 presents and discusses the results of outlier analysis.

2 Indicator Selection and Data Preprocessing

2.1 Indicator Selection

Maize growing is collectively affected by factors such as soil nutrition, moisture, temperature, light, carbon dioxide and so on [15]. Because its nature of low cold-tolerance, maize can grow and mature only when a certain effective accumulated temperature (EAT) is attained during its growing period. Its root system is generally shallow, so the plant is dependent on soil moisture. As a Carbon-4 (C4) plant, maize is a considerably more water-efficient crop than Carbon-3 (C3) plants like small grains, alfalfa and soybeans. Maize is most sensitive to drought at the time of silk emergence, when the flowers are ready for pollination. Because of manual fertilization, the request of soil nutrition is not as demanding. Consequently, soil moisture, temperature and light are the most crucial factors for maize growing. In this work, we select temperature, precipitation and elevation as the indicators for analyzing the differences of maize growing regions. In addition, we add maize yield as an auxiliary indicator. Taking into account the various growing periods in China, annual EAT, which is defined as the accumulated value of daily maximum temperature during the whole year as long as it is above 10 degree Celsius, is adopted instead of temperature. Similarly, precipitation is defined as the accumulated value of daily maximum rainfall during the whole year.

In order to achieve a more spatially coherent clustering result, we also add the coordinate of the center of each county as one of the clustering indicators.

2.2 Data Sources

Elevation is the mean of intersection result of the Digital Elevation Model (DEM) and administrative county region geographic data, whose scales are both 1:4,000,000. Coordinate data is extracted from the administrative county region geographic data by ArcMap. Precipitation and EAT data are derived from 671 meteorological observatories from 1951 to 2005. Spatial interpolation is applied to the precipitation and EAT of those 671 sites, to extract the data for the counties. Yield is the value of county mean yield divided by the county mean maize planting area from 1990 to 2005.

2.3 Data Refinement

Incomplete data and singular values of a few counties are found during the data verification process, i.e. maize mean yield is 0 or exceed 2000. Data cleaning is necessary to deal with those dubious cases; otherwise the clustering results would be impaired. We replace the singular data with the corresponding attribute mean of the contiguous regions. Notice that this process would not necessarily affect the outlier detection afterwards, since we only remove the data that is logically impossible to obtain.

2.4 Data Normalization

We perform min-max normalization on the original data so that the attribute data are scaled to fall within the same specified range. Let min A and max A denote the minimum and maximum values of an attribute A. Min-max normalization maps a value v of A to v’ in the range [0,1] by computing

$$v' = \frac{v - \min_A}{\max_A - \min_A}. \quad (1)$$

3 Proposed Spatial Clustering Algorithm

We propose a spatial clustering approach that consists of two steps. First, attributive clustering is applied on the indicators (with spatial information, refer to Section 1.1). We select K-means clustering algorithm in this step. Second, post-processing step of spatial contiguity adjustment is applied according to specific rules. This two-step algorithm results in clustered regions which are both contiguous in spatial relation and similar in attributes.

3.1 Attributive Clustering

We apply K-means algorithm [4] for attributive clustering. Euclidean distance was chosen to measure the similarity between two counties. Suppose entity A contains n attributes and entity B contains n attributes , the similarity of the two entities is given by:

$$S = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + (X_3 - Y_3)^2 + ... + (X_n - Y_n)^2} \quad . \qquad (2)$$

In this paper, the similarity measure specifies to:

$$S = \sqrt{\Delta r^2 + \Delta t^2 + \Delta e^2 + \Delta o^2 + \Delta x^2 + \Delta y^2} \qquad (3)$$

where △r, △t, △e, △o, △x and △y are the difference of precipitation, EAT, elevation, yield, x- and y- coordinate between two counties.

In K-means, the number of clusters needs to be specified as a parameter for the clustering process. According to the literature on regionalization of maize cultivation in China [16], we specify the number of clusters to be 7. K-means, by nature, may converge to different local minimum when the initialization is different. To obtain a result that is close to the global optimum, we repeated the K-means algorithm 100 times, each time with a different initialization, and select the run that leads to the minimum sum of variance. Three sets of repeated experiments show that the final result obtained is consistent with each other. Fig. 1 gives the attributive clustering result.

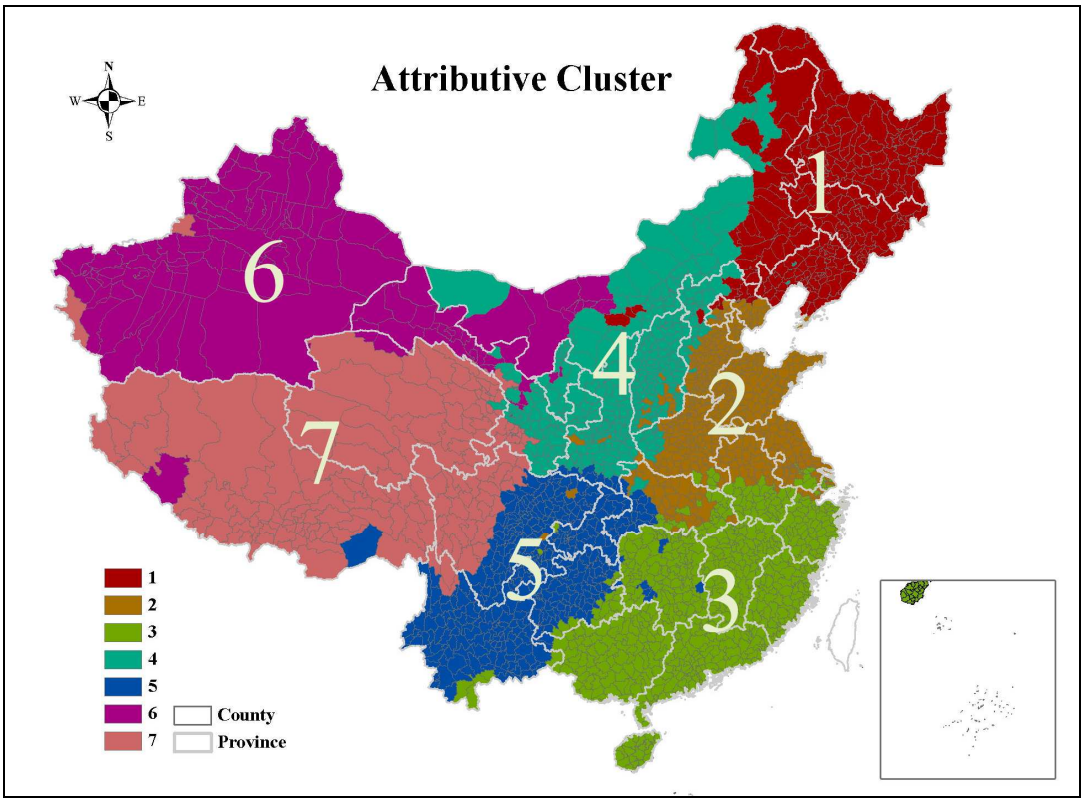

Fig. 1. Attributive clustering result.

(Note: We considered the islands be contiguous with its closest mainland region. For example, Hainan province is considered as contiguous with Guangdong province. (Same as below))

3.2 Spatial Contiguity Adjustment

From the first-step clustering result, we observe that, although on the whole there is strong spatial coherence in the result of the attributive clustering, there are several circumstances which do not meet the regionalization requirement. Spatial contiguity adjustment on some entities is necessary to make the spatial clustering satisfy the requirement that entities in the same cluster should be contiguous.

After fully considering the spatial contiguity requirement, we formulate several rules to adjust the result of attributive clustering, as follows:

Case 1: If the county is wholly surrounded by some cluster A, classify the county as in cluster A, as is shown in Fig. 2a and 2b.

Case 2: If the county is contiguous with several clusters, compute the differences between the single county and the contiguous clusters and classify it as in the cluster with the minimum difference, as exhibited in Fig. 2c and 2d.

Case 3: If the county is sandwiched between two areas of the same cluster, as demonstrated in Fig. 2e where the green cluster is cut off by the red one, find a route which would connect the green cluster with the smallest change. In the case shown in Fig. 2e and 2f, classifying the highlighted red region as green is the best way.

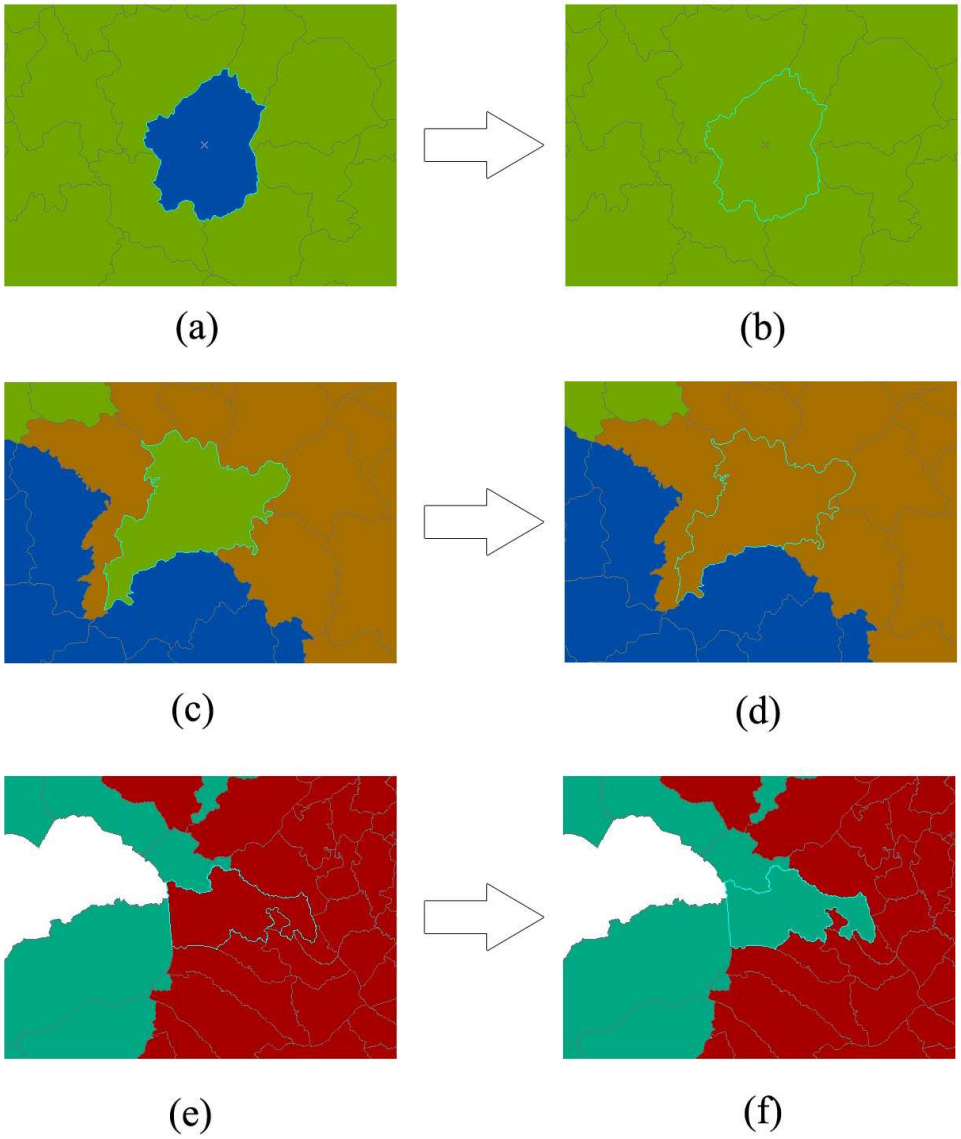

Fig. 2. Adjustment rules for spatial contiguity.

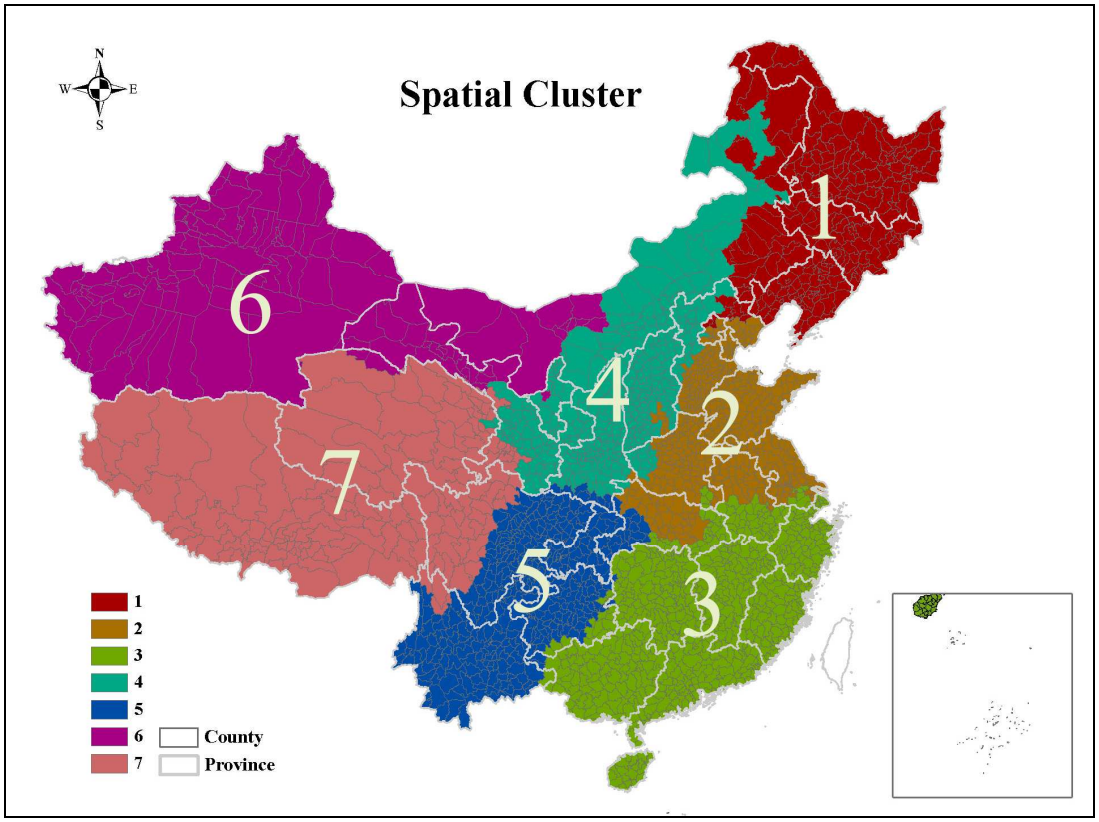

Fig. 3. The final spatial clustering result.

After all the adjustments, we gain spatial clustering a result which has the desired properties, i.e. entities within the same cluster are as similar as possible and entities in different clusters as different as possible, and entities of the same cluster are contiguous. Fig. 3 shows the final clustering result.

4 Outlier Detection

Spatial clustering aims to find the clusters of objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. However, very often, there exist data objects which do not comply with the general models of the data. Such sets of data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers of the data set. To a cluster, objects which are different from the remaining set in the cluster are defined as outliers of the cluster. In this work, outlier analysis can be used to provide supporting and assistant information, which can help identify the abnormal counties.

4.1 Dissimilarity Indicator

The criterion for selecting outliers should be consistent over all clusters. Consequently, each county was compared with the object the attributes of which was mean of all the counties in the corresponding cluster the county belongs to. The same with the measurement method adopted in attributive clustering, we adopted Euclidean distance to measure the difference between two entities in a cluster. However, unlike (2) coordinates are eliminated in the dissimilarity indicators, since the proximity of counties should not be an influential factor for detecting outliers. Hence, the dissimilarity is given by:

$$D=\sqrt{(r-\overline{r})^2+(t-\overline{t})^2+(e-\overline{e})^2+(o-\overline{o})^2} \quad (4)$$

where $\overline{r}$, $\overline{t}$, $\overline{e}$ and $\overline{o}$ are the mean of precipitation, EAT, elevation and yield of the cluster the data belongs to.

4.2 Outlier Extraction

We sort the dissimilarity value in descending order, and select the first 5% as the outliers. After computation, the counties with the dissimilarity value larger than the threshold 0.2576 were selected as the

outliers. Table 1 gives the number of outliers and their proportion in each cluster and Fig. 4 displays their distribution in a map.

4.3 Outlier Analysis

As indicated in Table 1, no county was regarded as an outlier in cluster 1 and 2, which means those clusters are highly consistent; only 1.42% of counties were determined as outliers in cluster 4, suggesting that cluster 4 is quite consistent as well; the large proportions of outliers to total counties in cluster 5, 6 and 7 worth some further discussions, as follows.

4.3.1 Influence of Each Indicator to Outlier Detection

To examine the influence of each indicator to the outlier detection, the same dissimilarity threshold (0.2576) was adopted to classify abnormal value of each indicator. In other words, if the value v of a indicator I of a county A meets the outlier condition, i.e., or where is mean of the indicator in a cluster), I is considered as the dominant factor which induces A be classified as an outlier. The result indicates that 76 outliers have a dominant factor (one of them has two dominant factors, precipitation and EAT), accounting for 64% of the total outliers. Detailed data is presented in Table 2.

TABLE 1

Statistics of Outliers in Each Cluster

cluster	1	2	3	4	5	6	7	∑
number of outliers	0	0	33	5	55	11	15	119
total of county	222	514	635	352	387	118	153	2381
proportion	0.00%	0.00%	5.20%	1.42%	14.21%	9.32%	9.80%	5.00%

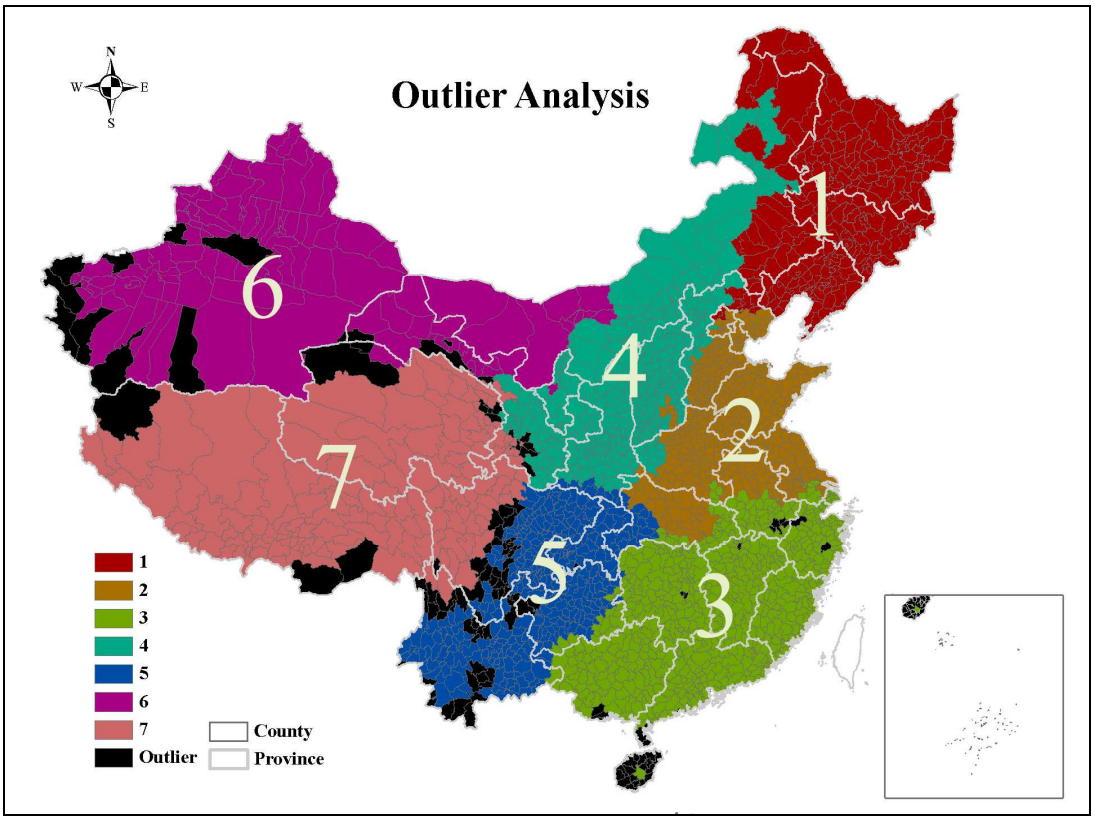

Fig. 4. Map display of distribution of outliers.

TABLE 2

The contribution of each indicator to outlier detection

Cluster	Elevation	EAT	Precipitation	Yield	∑
3	0	24	2	0	26
4	3	0	0	0	3
5	13	11	7	2	33
6	7	0	0	1	8
7	6	1	0	0	7
∑	29	36	9	3	77
proportion	38.16%	47.37%	11.84%	3.95%	100%

Note: Cluster 1 and 2 were not analyzed because there is no outlier in those clusters.

In examining the dominant factors in creating outliers, we find that, 36 outliers, which is nearly 50% of total outliers, has EAT as its dominant factor. We also find that for 38% the dominant factor is elevation. Hence, we conclude that EAT and elevation have strong influence in creating outliers.

In examining different regions cluster-by-cluster, we find that the numbers of outliers which have a dominant factor in cluster 3 and 5 are greater than in other clusters. The following analysis focuses on cluster 3 and 5.

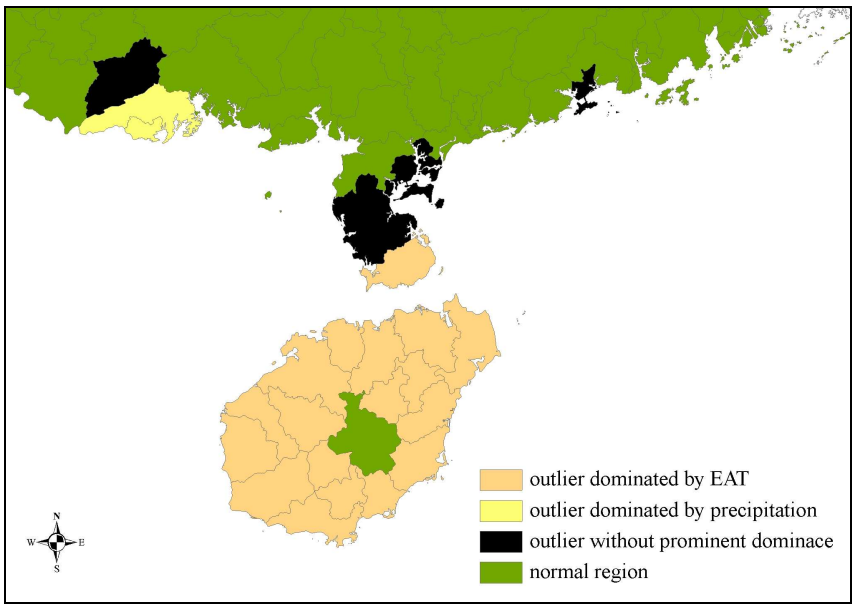

Fig. 5. Outliers of cluster 3.

As shown in fig. 5, in cluster 3, the majority counties in Hainan province were classified as outliers because of excessively high EAT, which is consistent with the reality. Hainan has a tropical moist monsoonal climate, thus its annual mean temperature is 23 to 25 degree Celsius. Except for the mountainous regions in the central part of the island, the daily mean temperature in Hainan is above 10 degrees Celsius, and the accumulated temperature during the growing season of the crops reaches eight thousand to nine thousand degree Celsius-days. Consequently, it is reasonable that majority of Hainan, apart from the central part, were categorized as outliers.

In the outliers of cluster 5, as showed in Fig. 6, a banded area in northwest is plotted as outlier because

of the high elevation. After overlaying with provincial boundary (shown in Fig. 7), we find that they belong to midland of Sichuan province and the northwest of Yunnan province. Because of the existing of Tibetan Plateau, hypsography rises in midland of Sichuan province. Northwest of Yunnan lies in Hengduan Mountain, which has mountain climate. Climate changes rapidly from hundred meters to several kilometers above the sea level. Hence, it is reasonable to classify these areas as outliers. Another outlier agglomeration is in the south of Yunnan Province, which is induced by the high EAT. This area belongs to Xishuangbanna and south of Puer, which is located in the south of Tropic of Cancer. It has a megathermal climate, which has high temperature throughout whole year.

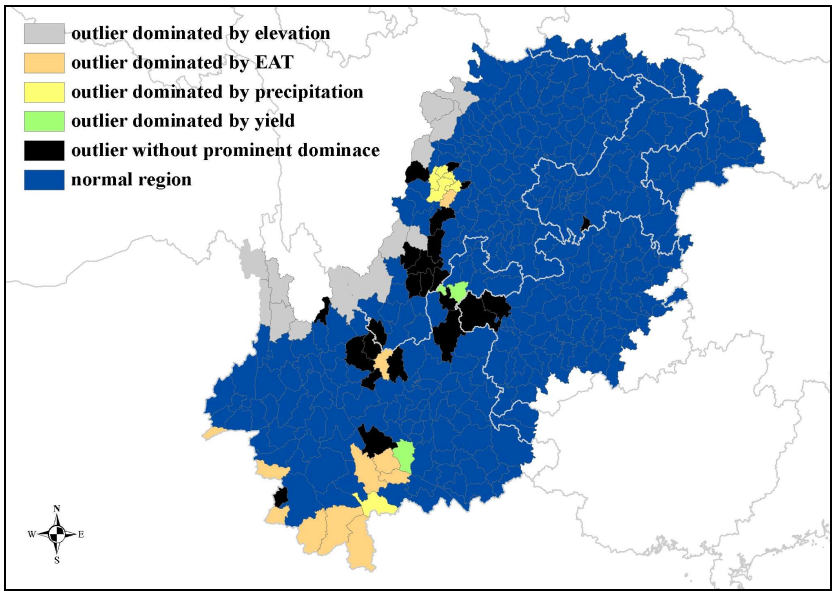

Fig. 6. Outliers of cluster 5.

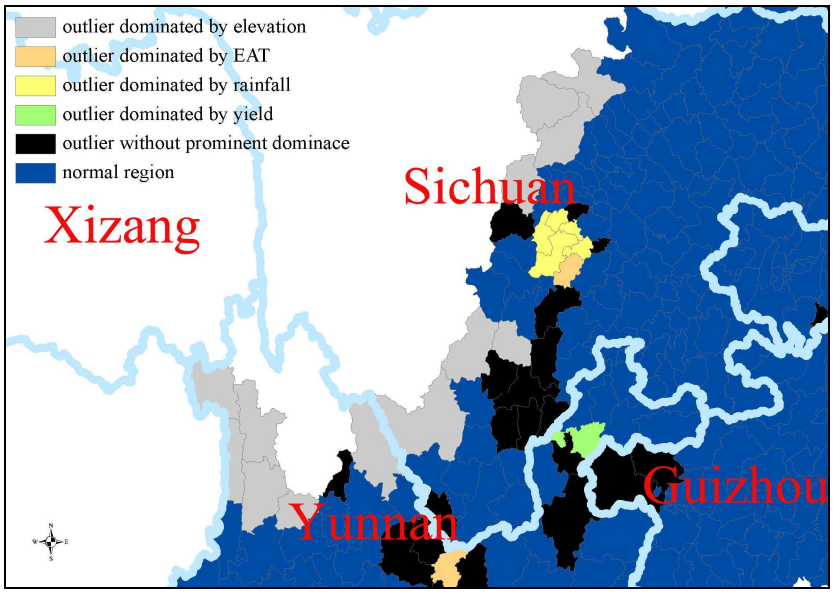

Fig. 7. Map display of outliers of cluster 5 overlaid with provincial boundary.

4.3.2 Influence of Adjustment of Spatial Adjacency to Outlier Detection

To examine the influence of Adjustment of Spatial Adjacency to outlier detection, we verify 66 adjusted counties and find that only 6 counties (account for

9%) are categorized as outliers. Therefore, we conclude that the spatial adjacency adjustment places little impact on the outlier detection.

5 Conclusion

In this paper, based on an approach of spatial clustering of continuous polygons, with the spatial adjacency adjustment as a post-processing step, we have gained a regionalization result of maize cultivation in China. We have also applied outlier analysis on each of the regions to identify the counties that show abnormal characters in maize cultivation, and further analyze the possible causes. We draw the following conclusions based on the results:

(1) The result of spatial clustering basically matches the conventional maize growing regions in China (as shown in Fig. 8), but also divides Region 1 into two regions, which is in accordance with the current new trial of maize cultivation regionalization applied in China [12]. The results prove the effectiveness of the proposed clustering approach.

Fig. 8. Maize growing region in China.

(2) The result of attributive clustering displays a strong spatial coherence as we expected and the number of entities which have been altered in the spatial adjacency adjustment accounts for 2.78% of all, which shows that taking into account the coordinates of counties during clustering is instructive and helpful.

(3) Only 9% of those entities were identified as outliers, which illustrates that the consistence of data in clusters is hardly dependent on adjustment. This shows the feasibility of the proposed spatial clustering algorithm.

This paper tries to select elevation, EAT, precipitation and yield as cluster indicators. If the indicator could be more elaborate and distributed appropriate weight, a more accurate result will be gained. Moreover, approaches of attributive clustering and rules of adjustment merit a closer and further exploration.

6 Acknowledgments

We would like to thank the Seed Industry Informationization Institute of China Agricultural University for providing all research data. This research is sponsored by the National Natural Science Foundation of China (40971055) and National Key Technology R&D Program under the Eleventh Five-Year Plan of P.R. China (2006BAD10A01).

References:

[1] Tong, P. Y. (1999). Regionalization of maize cultivation in China. Beijing, China: China Agricultural Technology Press (in Chinese).

[2] Liu, Y. H., Zheng, D., Ge, Q. S., & Wu, S. H. (2005). Problems on the research of comprehensive regionalization in China. Geographical Research, 24, 321-329 (in Chinese).

[3] Zheng, D., Ge, Q. S., & Zhang, X. Q. (2005). Regionalization in China: retrospect and prospect. Geographical Research, 24, 330-344 (in Chinese).

[4] Han, J. W., & Kamber, M. (2001). Data mining: concepts and techniques. San Francisco, USA: Morgan Kaufmann Publishers.

[5] Renata, I., & Ferenc, K. (2005) Clustering Techniques Utilized in Web Usage Mining. Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, pp237-242.

[6] G. Manco, R. Ortale, & D. Sacc. (2003) Similarity-based clustering of web transactions," in SAC '03: Proceedings of the 2003 ACM symposium on applied computing, pp. 1212–1216.

[7] Li, X. L., Chen, Y. Y., etc. (2008) ComGIS-based Decision Support System for Land-use Structure Optimization. Proceedings of the 13th WSEAS international conference on applied mathematics, pp. 297-302.

[8] Harvey, J. M., & Han, J. W. (2001). Geographic data mining and knowledge discovery. London, England: Taylor and Francis.
[9] Luc, D. R., & Arno, S. (2001). Principles of data mining and knowledge discovery. Berlin, Germany: Springer.
[10] Zhang Yan, Zhang Xiaodong, Zhu Dehai, Huangzhimin. (2007) A weighted-Voronoi Polygons algorithm which is based on integration of spatial and non-spatial attributes, for space segmentation. WSEAS Transactions on Information Science and Applications, 4 (5): 1019-1025
[11] Qian, S. Q., Chen, X. P., & Guo, J. M. (1994). Application of cluster analysis for dividing flue-cured tobacco growing regions. Journal of Anhui Agricultural University, 21, 21-25 (in Chinese).
[12] Wang, S. J., & Ni, C. J. (2008). Application of projection pursuit dynamic cluster model in regional partition of water resources in China. Water Resources Manage, 22, 1421-1429.
[13] Yuan, S. M., Wen, T. Y., Xin, X. M., & Xue, J. T. (2007). Application of K-means for the regionalization of Chinese Angelica introduction in the Linxia Hui Autonomous Prefecture. Journal of Gansu College of Traditional Chinese Medicine, 24, 51-54 (in Chinese)
[14] Yang, S. Q., & Yu, S. H. (1999). Application of cluster analysis for the regionalization of precipitation area in the Sichuan Basin. Journal of Sichuan Meteorology, 19, 38-39 (in Chinese).
[15] Yu, Z. W. (2003). Crop cultivation: North. Beijing, China: China Agricultural Press (in Chinese).
[16] National Agro-Tech Extension and Service Center. (2007). National trial report of maize variety in 2001-2007. Beijing, China: China Agricultural Science and Technology Press (in Chinese).