Spatial Layout of Multi-Environment Test Sites: A Case Study of Maize in Jilin Province

Zuliang Zhao 1,2,†, Liu Zhe 1,2, Xiaodong Zhang 1,2,*, Xuli Zan 1,2, Xiaochuang Yao 3, Sijia Wang 1,2, Sijing Ye 4, Shaoming Li 1 and Dehai Zhu 1,2

1 College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China; zlzhao@cau.edu.cn (Z.Z.); liuz@cau.edu.cn (L.Z.); zanxuli95@163.com (Xl.Z.); wangsji@cau.edu.cn (S.W.); lshaoming@sina.com (S.L.); zhudehai@263.net (D.Z.)
2 Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture, Beijing 100083, China
3 Satellite Data Technology Division Institute of Remote Sensing and Digital Earth, CAS, Beijing 100093, China; yaoxc@radi.ac.cn
4 State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China; yesj@bnu.edu.cn
* Correspondence: zhangxd@cau.edu.cn; Tel.: +86-1390-113-3526
† Current address: Department of Computer Science, The City College of New York, New York, NY, USA.

Received: 9 April 2018; Accepted: 26 April 2018; Published: 4 May 2018

Abstract: Variety regional tests based on multiple environments play a critical role in understanding the high yield and adaptability of new crop varieties. However, the current approach mainly depends on experience from breeding experts and is difficult to promote because of inconsistency between testing and actual situation. We propose a spatial layout method based on the existing systematic regional test network. First, the method of spatial clustering was used to cluster the planting environment. Then, we used spatial stratified sampling to determine the minimum number of test sites in each type of environment. Finally, combined with the factors such as the convenience of transportation and the planting area, we used spatial balance sampling to generate the layout of multi-environment test sites. We present a case study for maize in Jilin Province and show the utility of the method with an accuracy of about 94.5%. The experimental results showed that 66.7% of sites are located in the same county and the unbalanced layout of original sites is improved. Furthermore, we conclude that the set of operational technical ideas for carrying out the layout of multi-environment test sites based on crop varieties in this paper can be applied to future research.

Keywords: spatial clustering; spatial stratified sampling; Multi-environment; maize

1. Introduction

Variety regional test, as a key to new crop variety performance and market prospects, has an irreplaceable role in breeding [1]. Since 2000, United States constructed a regional test network based on hundreds of test sites to represent almost all types of planting environments [2]. China has also built a systematic regional test network [3–5]. To accurately assess each variety within 2–3 years, every test site must be highly representative of planting environments, which cover several elements such as weather, soil, terrain, biological factors, etc., called multi-environments. However, regional test results are still inconsistent with actual crop results. An important reason for this result is that neither the number nor the locations of test sites could adequately represent the multi-environments.

Clustering, a fundamental method in regionalization based on multi-environments, has been used in maize planting environments for different applications. To deal with the sparse data from observed...
stations such as meteorological data, previous studies can be divided into two major categories: (1) studies that focus on environments clustering of the site itself, for instance, ecological and climatic factors limiting maize production including drought [6–8], sunshine [9], insufficient accumulated temperature [10], high temperature heat damage [11], and terrain conditions [12]; and (2) studies that build index system for clustering based on transferring the observed station data mainly on calendar time into the regular grid data by spatial interpolation method [13–18]. In fact, the environmental conditions of phenophase have a greater impact on crop varieties.

The accurate prediction of new varieties would be increased by the accumulated data from the same test sites for years. Therefore, selecting suitable test sites is a prerequisite for successful testing. The factors affecting the selection of test sites are not only limited to multi-environments [19,20], but also include flat terrain, uniform fertility, good irrigation, drainage conditions, and convenient transportation [21]. Currently, the existing system mainly concerns representativeness, stability, and yield of every site. The study results from researchers whom mainly focus on the representativeness, stability, and area discrimination of every site and rarely consider the planting area [22] are difficult to apply.

Planting environmental factors, as the geographical objects, usually have a certain spatial correlation. The traditional sampling methods could not apply to planting environment. Spatial sampling algorithms that are very popular [23–25] are seldom used in test site layout. In some previous practical applications [26–29], sites were selected based on expert knowledge. However, as the theoretical research continues to deepen in the fields of spatial sampling algorithms and spatial autocorrelation algorithms, researchers began to use spatial sampling models to optimize the sampling results to ensure the accuracy and avoid the bias caused by the subjective judgment.

In this paper, we propose a three-stage spatial layout method: (1) based on meteorological data, soil nutrient data, and topographical data during the phenophase period, we clustered the planting environment by spatial clustering method; (2) we used spatial stratified sampling to determine the minimum number of test sites in each type of environment; and (3) combined with factors such as the convenience of transportation and the planting area, the layout of multi-environment test sites was constructed according to spatially balanced sampling method. We take maize in Jilin Province, one of the main maize producing areas, as a case study. The experimental results were compared with existing sites to verify the availability of our method.

2. Materials and Methods

2.1. Index System Construction

The selection of indicators is the fundamental step for environmental recognition. In this paper, we use indicators as a theoretical basis for the layout of maize variety regional test sites which are used to select new varieties that are not only suitable for planting in different environments but also stable and high yield. The first principle of selecting indicators is the ecological factors that have a great impact on maize production but are independent from each other. The limiting factors of maize production mainly include cultivation and planting techniques, ecological and climatic factors, seed variety, soil, and biological stress [30]. Here, we use the following four ecological and climatic factors to build the indicator system: let \( n, i \) denote the days and index in grown period.

1. Accumulated temperature (AT) which refers to the accumulation daily average temperature (\( t \)) from sowing to maturity is formulated as:

\[
AT = \sum_{i=1}^{n} t_i \quad \text{subject to} \quad t_i \geq 10 \, ^{\circ}C
\]
2. Accumulated Precipitation (AP) which refers to accumulation of precipitation during whole grown period is formulated as:

\[ AP = \sum_{i}^{n} P_i \]  

(2)

\( P_i \) refers to daily \( i \) precipitation (units:mm)

3. Cumulative Sunshine Hours (CSH) which refers to accumulation of sunshine hours during whole grown period is formulated as:

\[ CSH = \sum_{i}^{n} S_i \]  

(3)

4. Elevation and Slope

2.2. Data Pre-Processing

In this study, meteorological data were provided by National Meteorological Center of China. The phenological data of spring maize and summer maize were obtained from China crop growth and farmland soil moisture data were provided by the National Meteorological Information Center of China and the China Ministry of Agriculture. Meanwhile, we collected some survey data from Agro-Seed Industry Companies. The national geography data used in this study were 1:4 million scale provincial and county administrative division vector data. The slope data came from the DRM (Shuttle Radar Topography Mission) 90 m resolution DEM data [31]. The distribution of test sites and weather stations in Jilin Province [32] is shown in Figure 1.

![Figure 1. Test sites and weather stations in Jilin Province.](image-url)
We used the Global Moaran’s I index and Z points to measure the spatial autocorrelation of each indicators. Positive and negative Maran’s I index values represent the corresponding positive and negative correlations of indicators. If Z points > 1.96, the spatial object is aggregated. If Z points < −1.96, the spatial object is decentralized [33]. We used the normal QQ plot to see if the data are normally distributed, and normalized the data using Equation (4).

\[ Z = \frac{x - \text{min}(X)}{\text{max}(X) - \text{min}(X)} \] (4)

where \( Z \) denotes the normalized value of each pixel (\( x \)), and \( \text{min}(X) \) and \( \text{max}(X) \) denote the minimum and maximum values of all pixels (\( X \)) before normalization, respectively.

### 2.3. Spatial Clustering

We proposed an integrated clustering algorithm for spatial attributes based on ISODData method [34,35], where the implementation is in three phases. (1) Cluster pedigree maps: \( R^2 \) and semi-\( R^2 \), are used to evaluate the clustering results and determine the number of clusters. (2) ISODData clustering algorithm is used for the planting environment. (3) Spatial continuity adjustments are made according to spatial adjustment rules. Partitions obtained have as many differences between classes as possible, small intraclass differences, and spatial continuity.

Four criteria for determining the number of clusters based on pedigree maps are as follows: (1) The distance between the centres of gravity are as far as possible. (2) The number of classes must meet practical purposes. (3) Practical purposes lead to the number of clusters. (4) The results obtained by different clustering methods should have the same class. Assume that sample size of \( n \) is divided into \( k \) categories, as \( C_1, C_2, \ldots, C_k \). \( n_i \) denotes the number of samples in Class \( C_i \). Let \( \bar{X}_i, \bar{Y}_i \) denote the center of gravity and the \( i \)-th (\( i = 1, \ldots, n_i \)) sample of \( C_i \). \( R^2_k \) is defined as follows:

\[ R^2_k = 1 - \frac{\sum_{i=1}^{n_i} W_i (X_i - \bar{X}_i)^2}{T} \] (5)

where \( W_i = \sum_{i=1}^{n_i} (X_i - \bar{X}_i)^2 \) and \( T = \sum_{i=1}^{n_i} \sum_{i=1}^{n_i} (X_i - \bar{X}_i)^2 \). We use \( R^2_k \) to evaluate the performance of the clustering with \( k \) clusters. The larger \( R^2_k \), the better performance based on \( k \) clusters. We also choose the semi-\( R^2 \) in this paper:

\[ \text{semi-} R^2_k = R^2_{k+1} - R^2_k \] (6)

A larger semi-\( R^2_k \) means a better performance of \( K+1 \) clusters.

Two spatial data adjustment rules for the raster data are defined as follows. Scenario 1: Other clusters distribute sporadic in a certain cluster. We use area threshold to determine whether the sporadic areas are retained. Scenario 2: the areas neighbor multiple clusters. Here, we calculate the difference between this area and all neighboring clusters and merge with the nearest one. The difference value \( D \) is defined as:

\[ D = \sqrt{\frac{\sum_{i=1}^{n_i} (X_i - \bar{Y}_i)^2}{n}} \] (7)

where \( \bar{X}_i \) denotes the mean of the \( i \)-th attribute of class of this area and \( \bar{Y}_i \) denotes one of the near cluster. \( n \) is number of attributes of one class.

### 2.4. Sample Strategy

We used a spatial sampling model proposed by Zhao et al. [32]; in this study, the relationship between the number of test sites \( x \) and sampling accuracy is defined as:

\[ x = \frac{(\sum W_h S_h \sqrt{C_h} \sum \frac{W_h S_h}{\sqrt{Y_h}})}{y + \frac{1}{n} \sum W_h S_h^2} \] (8)
The number of samples in each layer is calculated as the following formula:

\[ n_h = \frac{W_h S_h / \sqrt{C_h}}{\sum W_h S_h / \sqrt{C_h}} \times n \]  

where \( N \) is the total number of grid samples, \( n \) is the number of test sites, \( h \) is the planting environment class, \( N_h \) is the number of samples in the \( h \) type of planting environments, \( W_h = \frac{N_h}{N} \) is the weight of \( h \) type of planting environments, \( S_h \) is true standard deviation of the \( h \) type of planting environments, and \( C_h \) is the cost of investigating a single sample of a planting environment.

### 3. Results

#### 3.1. Data Processing

Moran’s I index of all three indicators is positive, which means all of them have spatial autocorrelation. The Z points are all greater than 1.96, indicating that the spatial distributions of the factors are clustered and the reliability is high (Table 1). Hence, we use statistical spatial interpolation for all three indicators.

**Table 1.** The Moran’s I index, Variance, and Z points of the three indicators.

<table>
<thead>
<tr>
<th>Indicators</th>
<th>Moran I</th>
<th>Variance</th>
<th>Z points</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>0.531548</td>
<td>0.006411</td>
<td>6.951109</td>
</tr>
<tr>
<td>AP</td>
<td>0.656345</td>
<td>0.006324</td>
<td>8.567889</td>
</tr>
<tr>
<td>CSH</td>
<td>0.685310</td>
<td>0.006488</td>
<td>8.818582</td>
</tr>
</tbody>
</table>

**Figure 2.** The normal plots for AT, AP and CSU of the whole growing period: (a) AT; (b) AP; and (c) CSH.

It can be seen in Figure 2 that all three factors are close to the normal distribution and the mathematical exceptions are unknown. Thus, we used the Ordinary Kriging method to interpolate AP and CSH. As the temperature is reduced by 0.6 °C for each 1000 m increased in elevation, we used Cokriging method to consider elevation for AT. The normal transformation parameters were not set and the grid resolution is 5000 m (Figure 3).
3.2. Multi-Environments Clustering

We divided the planting environments into 2–9 categories, and calculated the $R^2$ to evaluate the effect of clustering. The $R^2$ and semi-$R^2$ for different numbers of classes are shown in Table 2.

**Table 2.** $R^2$ and semi-$R^2$ for different numbers of clusters.

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>$R^2$</th>
<th>semi-$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>0.926</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.911</td>
<td>0.015</td>
</tr>
<tr>
<td>7</td>
<td>0.904</td>
<td>0.007</td>
</tr>
<tr>
<td>6</td>
<td>0.891</td>
<td>0.013</td>
</tr>
<tr>
<td>5</td>
<td>0.880</td>
<td>0.009</td>
</tr>
<tr>
<td>4</td>
<td>0.826</td>
<td>0.054</td>
</tr>
<tr>
<td>3</td>
<td>0.787</td>
<td>0.039</td>
</tr>
<tr>
<td>2</td>
<td>0.663</td>
<td>0.124</td>
</tr>
</tbody>
</table>

We found $R^2$ are all around 0.9 with no significant difference using more than five clusters. Furthermore, it is largest at four clusters, thus we selected five as the number of clusters to divide the planting environments (Figure 4).
We used spatial continuity adjustment rules in the planting environment of Jilin Province as a case study. Two methods are used to adjust planting environment (Figure 5).

Figure 4. Preliminary classification results.

Figure 5. Spatial continuity adjustment: (a) Sporadic distribution of other class (red middle) in a certain class (green background). They would be adjusted directly into the adjacent major categories. (b) A region (red upper) is adjacent to more than one class (purple upper and green background) at the same time. Thus, the two types of minimum values are classified into one. The colors filled in the grids represent the same clusters as in Figure 4.
Maize planting environmental divisions is shown in Figure 6.

![Figure 6. Maize planting environmental divisions.](image)

As a result, combining geography and actual conditions, the planting environment in Jilin Province is divided into five districts: High Altitude Forest area of the Changbai Mountain (HAFCM), Low Altitude Forest area of the Changbai Mountain (LAFCM), Eastern part of the mountain Agriculture and Forestry Area (EAF), Central Plains Agricultural (CPA), and Western plain Agricultural and Pastoral Area (WAP).

### 3.3. Test Sites Layout

The layout of test sites based on multi-environments in Jilin Province mainly considers the following three issues: (1) the minimum number of test sites; (2) the number of sites for each type of planting environment; and (3) the site locations. Test sites layout consists of four steps:

**Step 1.** In Jilin Province, in addition to the site’s ability to fully represent the different regional planting environment, the following factors should be considered: the distance to roads, and the total planted area. We use the national and provincial road data to calculate the degree of convenience for traffic.

**Step 2.** Using Equation (8), we can conclude that at least 25 sites should be deployed to meet the sampling accuracy requirements (error = 0.05). In this study, we also considered that the cost of single-sample surveys of all types is equal. According to Equation (9), we obtained the number of each type: 

\[ N_{(HAFCM)} = 5, N_{(CPA)} = 5, N_{(LAFCM)} = 4, N_{(WAP)} = 5, \text{ and } N_{(EAF)} = 6. \]

**Step 3.** A probability grid, which was used for site location, was constructed (Figure 7) based on following three factors: (1) representation of the planting environment (pdist): the distance from the sample to the cluster center; (2) planting area (area); and (3) road distance (roaddist). Based on the expert knowledge, the weights of pdist, area, and roaddist in Jilin Province were set to: 

\[ w_{(pdist)} = 0.1, w_{(area)} = 0.8, \text{ and } w_{(roaddist)} = 0.1. \]

The calculation formula for the probability raster (prb) is:

\[
prb = \left( \frac{pdist}{max(pdis)} \times w_{(pdist)} + \frac{area}{max(area)} \times w_{(area)} + [1 - \frac{roaddist}{max(roaddist)}] \times w_{(roaddist)} \right) \times slope
\]
Step 4. We used spatially balanced sampling to set up a testing site for each type of planting environment. (Figure 8)

The sample relative accuracy calculated using Equation (8) is 94.5% when the number of test sites \( x = 25 \).
4. Discussion

4.1. Clustering Attribute Statistics

The statistics of five types of planting environment are shown in Table 3.

<table>
<thead>
<tr>
<th>Type</th>
<th>Range AT</th>
<th>Mean AT</th>
<th>Range AP</th>
<th>Mean AP</th>
<th>Range CSH</th>
<th>Mean CSH</th>
<th>Range Elevation</th>
<th>Mean Elevation</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAFCM</td>
<td>1950–2500</td>
<td>2300</td>
<td>180–300</td>
<td>240</td>
<td>900–970</td>
<td>926</td>
<td>2–1481</td>
<td>600</td>
</tr>
<tr>
<td>HAFCM</td>
<td>2040–2500</td>
<td>2250</td>
<td>250–360</td>
<td>300</td>
<td>900–1000</td>
<td>953</td>
<td>330–2670</td>
<td>860</td>
</tr>
<tr>
<td>EAF</td>
<td>2330–2770</td>
<td>2570</td>
<td>280–430</td>
<td>340</td>
<td>920–1050</td>
<td>980</td>
<td>90–1500</td>
<td>500</td>
</tr>
<tr>
<td>CPA</td>
<td>2600–2900</td>
<td>2740</td>
<td>240–330</td>
<td>280</td>
<td>1030–1180</td>
<td>1110</td>
<td>130–850</td>
<td>240</td>
</tr>
</tbody>
</table>

In HAFCM, the terrain fluctuated greatly, area of maize is less, and it is prone to frost damage because of the earlier frost period. In contract, AP and CSH of LAFCM are relatively small. EAF, a semi-mountainous valley, is generally basin and plain, where the arable land is relatively large and suitable for agricultural development. The climate of this area is mild, e.g. AT is between 2330 °C and 2770 °C with abundant AP. CPA is the main maize belt in Jilin Province. Most of the cultivated land is concentrated and contiguous, which is suitable for mechanized farming. It has abundant photothermal resources, for instance, AT is between 2600 °C and 2900 °C and AP is between 240 m and 330 mm.

4.2. Planting Environmental Representation

We clustered each type of planting environment using the same clustering method, and compared them with the test site layout results.

Test sites proposed in this paper cover 19 of 25 sub-clusters (Figure 9). However, the number of test sites is still unbalanced. Two reasons may cause this situation: (1) the randomness of spatially-balanced sampling; and (2) the planting areas in sub-clusters are too small.
4.3. Comparison Number of Test Sites

The number of original test sites in Jilin province is 26 (Table 4). The spatial distribution of the existing test sites is not well-balanced. For example, in CPA, the number of test sites is 13 which is about half of the total sites. The total established test sites number of HAFCM and LAFCM cannot reflect actual complexity, comparing with nine test sites from our method. In addition, EAF and WAP also need to add two sites according to our method.

Table 4. Number of test sites.

<table>
<thead>
<tr>
<th>Type</th>
<th>Established Sites</th>
<th>Our Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAFCM</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>LAFCM</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>EAF</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>CPA</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>WAP</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

5. Conclusions

To tackle with the problems of spare data from multi-environments and inconsistence between regional tests and actual promotion, we propose a spatial layout method that includes the following two novel features: (1) It constructs a clustering index system of planting environment with test site layout as the application purpose. (2) It deduces the appropriate spatial distribution of test points in different cluster by integrating the complexity of each planting environment type. The experiment was run in Jilin Province to simulate the layout of maize variety testing sites. The results show the proposed methods can not only meet requirements for quantity measurement and spatial distribution of test sites, but also provides a set of operational technical ideas for the layout of multi-environment test sites for crop varieties.

Author Contributions: Z.Z., L.Z. and Xd.Z. conceived and designed the experiments; S.W. and Xl.Z. performed the experiments; Z.Z. and L.Z. analyzed the data; and X.Y., S.Y., S.L. and D.Z. contributed analysis tools. All authors contributed to writing and reviewing the manuscript. All authors read and approved the final manuscript.

Acknowledgments: This paper is supported by the “Foundation for Key Program of Beijing” (Grant No. D171100002317002) and the China Scholarship Council (CSC) (Grant No. 201606350174).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

- AT Accumulated temperature
- AP Accumulated Precipitation
- CSH Cumulative Sunshine Hours
- LD linear dichroism
- HAFCM High Altitude Forest area of the Changbai Mountain
- LAFCM Low Altitude Forest area of the Changbai Mountain
- EAF Eastern part of the mountain Agriculture and Forestry Area
- CPA Central Plains Agricultural
- WAP Western plain Agriculture and Pastoral Area

References


15. Wang, D.; Li, G.; Mo, Y.; Cai, M.; Bian, X. Effect of Planting Date on Accumulated Temperature and Maize Growth under Mulched Drip Irrigation in a Middle-Latitude Area with Frequent Chilling Injury. Sustainability 2017, 9, 1500. [CrossRef]


