ANOVA for Spatial Data after Filtering out the Spatial Autocorrelation

Yumin Chen[1, a], Yan Lu*[1, b], Jiang Zhou[1, c] and Mo Cheng[1, d]

[1]School of Resource and Environment Science, Wuhan University, 129 Luoyu Road, Wuhan, 430079, China

[a]ymchen@whu.deu.cn, [b]lu_giser@outlook.com, [c]giser_zhou@outlook.com, [d]chengmof1@gmail.com

Keyword: ANOVA, spatial data, GIS, spatial statistics

Abstract. Spatial data contains a lot of spatial information. Geographic Information System (GIS) provides methods to extract information and support decision. The difference of geographic data can be found out by analysis of variance (ANOVA) that is a method of spatial statistics, which is a part of GIS. In this paper, we study how to do ANOVA for spatial data after filtering out the spatial autocorrelation, in order to discover the differences between spatial data. Through the experiment of ANOVA for spatial data of some counties of China, the difference of the total powers of agriculture machine of different regions which has not the same level per capita GDP is significant. This information of difference can be provided to the decision-making department for making decision.

Introduction

About 60% of all information we contact nowadays is geospatially referenced [1]. And the most effective way to analyze the spatial data to get information is using GIS. Spatial statistics uses the statistical knowledge to analyze spatial data for deeper and more detailed information.

Recently, an application of ANOVA to German stock for testing the mean volatilities by Jaechoul Lee [8], Jong Oh Choi using ANOVA to do the work of uncertainty evaluation [10]. Jan Gertheiss do ANOVA for ordered data to testing differentially expressed genes with ordinal phenotypes [4]. In these applications, the data which used to ANOVA isn’t consider the spatial autocorrelation, so the data may be not independent. In such a situation, ANOVA can’t be applied directly.

This paper will provide a method to analyze the spatial data with ANOVA, which aim at analyze whether there is a significant difference on a certain spatial attribute between different groups or different regions [3]. Significant differences between spatial objects will be presented directly to help make decisions. In this paper, the experiment is carried out in the open-source software R. And the key of the experiment include three parts: spatial autocorrelation analysis, ANOVA of spatial data, and the basic condition test of ANOVA [7] [8].

Methodology

The method used for this study is composed of three parts, namely, spatial autocorrelation analysis, ANOVA for spatial data and results analysis and evaluation.

1) Spatial Autocorrelation Analysis

Tobler’s First Law [2] is the best summarized of spatial autocorrelation. That is, the strong effect that nearby areas have on each other versus the relatively weak influence of areas further away with the implication that near spatial units are similar to one another [5]. The purpose of analyzing spatial autocorrelation is to eliminate the effects of its nearby spatial units, so as to meet one of the basic conditions of ANOVA: independence. To filter out the spatial autocorrelation is to convert the spatial data into ordinary data, that is, the data are independent of each other.

Spatial Error Model (SEM) is used to analyze the spatial autocorrelation, before this, a spatial contiguity matrix is necessary. Spatial contiguity matrix is the matrix that expresses the contiguity of any two entities in space. Spatial contiguity matrix is divided into topology-based contiguity matrix and distance-based contiguity matrix. In my experiment, topology-based contiguity matrix is chose. If one entity is connected with another entity, the corresponding in the matrix is assigned the value 1, otherwise assigned the value 0.

SEM is a regression model that analyze the correlation between two variables. The traditional regression model is no longer applicable when the variables is dependent. After eliminating the spatial autocorrelation, the linear regression model is working. The equation of SEM [6] is following Eq. 1

$$y = \alpha x + (I - \rho W)^{-1}\varepsilon \qquad (1)$$

Where x, y are the variables in model, ρ is the coefficient of spatial autocorrelation，W is the topology-based contiguity matrix, ε is the error of regressiom model, α is the constant coefficient of regression model.

With the premise that existence of the SEM, the value of ρ represents the degree of spatial autocorrelation. Then to filter out the spatial autocorrelation can convert the spatial into ordinary data. The formula is following Eq. 2:

$$\bar{y} = y * (I - \rho W) \qquad (2)$$

2) ANOVA for spatial data

ANOVA, also called F test, is used to analyze the significant difference between different groups of data. ANOVA is also a kind of regression. But the ordinary regression analysis is the relationship between two or more variables while the ANOVA is the relationship between one variable and one or more factor variables (the levels of a discrete covariate), or the relationship between several variables and one factor variable. ANOVA is divided into different methods according to the numbers of variables and factor variables. The method used in this paper is one way ANOVA, that is, only one variable and one factor variable [4].

The key of ANOVA is to analyze whether the error between different groups or the random error between data in the same group is more significant. If there is a significant difference between different groups, that means obvious differences between the two groups. Otherwise, it means the difference between the groups is insignificant. ANOVA can effectively point out whether one attribute is different in different regions [9]. Similarly, it can also be grouped by a certain attribute grade. For example, grouping data by different GDP per capita, we can analyze whether there is significant difference between per capita water resources in different groups.

Prior to the ANOVA, the null hypothesis should be put forward to analyze the problem. It’s often represented as “There is no significant difference of a certain index in the space under the groups of different regions or different levels. The final result is the acceptance or rejection of the null hypothesis.

There are three fundamental conditions of ANOVA: independence, normality, homoscedasticity. The above analysis of spatial autocorrelation of the data and the elimination of spatial autocorrelation in Section II, ensure the independence of the data. And there are many ways to test the normality and homoscedasticity of data. In this experiment, we use QQ-Plot to test the normality according to the image. If the data fall within the 95% confidence intervals (CI), the data will satisfy the normality assumption. The test of the homoscedasticity is a test of residual after ANOVA, we use Levene test for homogeneity test. If the residual satisfy the homoscedasticity, the original data will be also satisfied. The results of ANOVA are often presented in the form of a table, as the Table 1.

Table 1: ANOVA Formula Table

	Sums of squared residuals	Degrees of freedom	Means of squared residuals	F
Treatment (variability between groups)	$SS_{Treatment} = SS_{Total} - SS_{Error}$	$df_{Treatment}$: N − 1	$MS_{Treatment}$: $\frac{SS_{Treatment}}{df_{Treatment}}$	$F = \frac{MS_{Treatment}}{MS_{Error}}$
Error (variability within groups)	SS_{Total}	df_{Error}: n − N	MS_{Error}: $\frac{SS_{Error}}{df_{Error}}$	

Where $SS_{Total} = \sum_{i=1}^{n}(X_i - \overline{X_{GM}})$, $SS_{Error} = \sum_{i=1}^{n}(X_i - \overline{X_{Gi}})$, X_i is the data in the data set, $\overline{X_{GM}}$ is the mean of all data, $\overline{X_{Gi}}$ is the mean of one group, N is the number of the groups, n is the total number of data.

3) Results analysis and evaluation

P-value is the probability that Pr(>F), P-value is usually used to explain the significant difference instead of using F-value directly. At the 95% CI, while the P-value (> F) is bigger than 0.05, we cannot reject the null hypothesis. That means, there is no significant differences between a certain spatial index in the groups of different regions or different grades. But when P-value is smaller than 0.05, we reject the null hypothesis, and that means there is a certain index having significant differences in the groups of different regions or different grades. In some applications, such as ANOVA for per capita GDP and the total powers of agriculture machine, the total powers of agriculture machine in the region whose per capita GDP is low level is also small.

Experiment

1) Study area and data

In the experiment, some counties in China are selected as the study area. There are 83 counties in the area in total, and the total powers of agriculture machine is chose as the variable of ANOVA, while GDP per capita is chose as the factor of ANOVA.

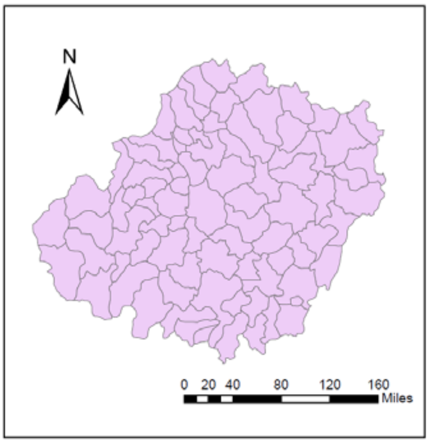

Figure 1: The map of study area

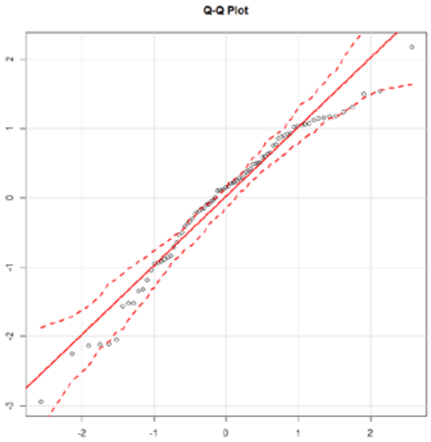

Figure 2: Normality Test

2) Experimental procedure and result

We do the experiment in the software R. The first step of the experiment is the analysis of spatial autocorrelation. The result of the analysis of spatial autocorrelation is that: the coefficient of spatial autocorrelation of the total powers of agriculture machine is 0.271, after eliminating the spatial autocorrelation, the coefficient decreased to -0.0068. The result of normality test is appeared in Figure 2, if the point of all of data is within the range of 95% CI, the data is normal. The Shapiro-Wilk normality test shows the data is normal because the P-value is 0.1165 (bigger than 0.05). So the data of our experiment is normal.

When we do homogeneity test of variance, the Levene Test is available. A P-value can be got as the result of the homogeneity test. If the p-value > 0.05, the conclusion that the data is homogeneous can be got easily. The Table 2 shows the result that the data is homogeneous. ANOVA can be in progress if the data is independent, normal, homogeneous, the result is appeared in the Table 3, because of the Pr(>F) < 0.05, we can conclude the difference of overall levels is significant. That is, the difference of the total powers of agriculture machine of different regions which has not the same level per capita GDP is significant.

Table 2: Levene's Test for Homogeneity of Variance

	Df	F value	Pr(>F)
group	4	0.3755	0.8255
	78		

Table 3: ANOVA Table

	Df	SS	MS	F value	Pr(>F)
x	4	11.70	2.9244	4.641	0.00206
Residuals	78	49.15	0.6301		

Discussion

The result is reasonable, because it’s obvious that the district which has a poor economy has a small total power of agriculture machine. The main function of the GIS is to extract the information contained in the geographic data, which is invisible and in deep level, and to use this information to help decision-making. Applying the tools like spatial statistics to GIS can mine useful information more effectively. ANOVA is one of the effective tools to analyze the difference. The results we got from the experiment, whether the difference is significant and which groups are significantly different, can reveal the difference from both micro and macro layers, can display directly the difference of which areas or which groups of an indicator (such as vegetation coverage, per capita water resources, etc.). So the results can provide decision support for the government departments.

Conclusion

In this paper, a method that applies ANOVA to spatial data is presented, and can be used to analyze the difference. Firstly, the spatial autocorrelation is introduced briefly. Then we ensure the independence, normality, and homoscedasticity of the data around the basic conditions of ANOVA. Finally, we analyze the difference of the powers of agriculture machine between several districts which has different per capita GDP levels. The result is significantly, That is, the difference of the total powers of agriculture machine of different regions which has not the same level per capita GDP is significant. ANOVA is a very common tool, but it is rare to handle geographic data which contain spatial autocorrelation.

The method of ANOVA for spatial data mentioned in the paper can be widely used in geographic data, including some location-related economic data and many Internet data which contain position information. Information extraction and decision supporting are the ultimate goal of the ANOVA for spatial data. And in fact, it will surely make it.

Acknowledgements

The research was supported by grants from National High Technology Research and Development Program of China (Project No. 2013AA122302, 2013AA122301), funded by Open Research Fund Program of Shenzhen Key Laboratory of Spatial Smart Sensing and Services (Shenzhen University).

Reference

[1] Hahmann, S., & Burghardt, D. (2013). How much information is geospatially referenced? Networks and cognition. International Journal of Geographical Information Science, 27(6), 1171-1189.

[2] Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic geography, 234-240.

[3] Iversen, Gudmund R., and Mary Gergen. Statistics: the conceptual approach. Springer Science & Business Media, 2012, 473-502.

[4] Gertheiss, J. (2014). ANOVA for Factors with Ordered Levels. Journal of Agricultural, Biological, and Environmental Statistics, 19(2), 258-277.

[5] Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. Geographical analysis, 27(4), 286-306.

[6] Kazar, B. M., & Celik, M. (2012). Spatial AutoRegression (SAR) Model: Parameter Estimation Techniques. Springer Science & Business Media, 7-17.

[7] Tarlow, K. R. (2015). Teaching principles of inference with ANOVA. Teaching Statistics.

[8] Lee, J., & Ko, K. (2007). One ‐ way analysis of variance with long memory errors and its application to stock return data. Applied Stochastic Models in Business and Industry, 23(6), 493-502.

[9] Bondell, H. D., & Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. Biometrics, 65(1), 169-177.

[10] Choi, J. O., Nam, G. H., & Kim, B. J. A thought on uncertainty evaluation using ANOVA. Accreditation and Quality Assurance, 1-4.