Computers and Electronics in Agriculture 148 (2018) 272–279

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Error analysis and correction of spatialization of crop yield in China – Different variables scales, partitioning schemes and error correction methods

Guangxing Ji[a], Li Tian[a], Jincai Zhao[a], Yanlin Yue[a], Zheng Wang[b,*], Xu Yang[a]

[a] Key Laboratory of Geographic Information Science, Ministry of Education, East China Normal University, Shanghai 200241, China
[b] Institute of Science and Development, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Keywords:
Crop yield
Variables scales
Partitioning schemes
Error correction methods
Multiple variable regression
Spatialization
China

ABSTRACT

Spatialization of crop yield is beneficial to comprehensive analysis between interdisciplinary data. Multivariable linear regression models are often applied to spatialization of attribute data. The variables scales and the partitioning of China should be considered when the model is constructed. Different variables scales and partitioning schemes will inevitably results in different spatialization errors. Spatialization errors can be reduced by error correction methods. Different methods have different influence on the accuracy of crop yield spatialization. In this study, three variables scales were selected including prefectural scale, county scale and grid cell (1 km × 1 km). Five partitioning schemes (no partition of China, 7 regions of China, 9 regions of China, 10 regions of China, partitions of China by province) were considered. A total of 28 kinds of multivariable linear regression models were constructed with area of different types of farmland as independent variables, crop yields as dependent variables. Then, seven kinds of error correction methods were used to correct crop yield spatialization results. Three error evaluation indicators were selected to investigate the influence of different variables scales, partitioning schemes and error correction methods on the precision of spatialization results. The conclusions can be drawn as follows: (a) Nine models with intercept based on variables at regional scale could not be used to spatialize crop yield, while the others can be used for spatialization of crop yield. (b) The precision of the spatialization result based on the model without intercept is higher than that based on the model with intercept. (c) For models without intercept, precision of spatialization results increased first and then decreased with the refinement of partitioning scheme. (d) For models without intercept, the precision of spatialization results improved with scaling down of the variables scale from prefectural scale to county scale and grid scale. (e) Among the seven kinds of error correction methods, average correction method, weight coefficient correction methodII and weight coefficient correction method III can't be used to correct initial spatialization results. (f) Proportional coefficient correction method, weight coefficient correction methodI, weight coefficient correction method IV and weight coefficient correction method V can be used to correct initial results of spatialization. (g) The precisions of corrected spatialization products based on error correction methods, which can improve the precision of initial spatialization products, are very closely. This research made up for the deficiency of spatial error analysis of crop yield, explored the relationship between different sample scales and partitioning schemes and spatial error, compared the pros and cons of different error correction methods. Meanwhile, it also provided valuable information for other types of social and economic statistical data.

1. Introduction

Given a backdrop of global environmental dynamism and climate change, traditional geo-ecological processes have undergone drastic changes over the past few decades. The geographical processes are no longer simple natural processes, and the researches of ecological processes also are no longer confined to the dynamics and development in ecosystem. The integration and intersection of multiple disciplines is becoming an important characteristic of modern geo-ecological processes (Fu et al., 2006).

It is an important symbol of the combination of human activities and geo-ecological processes to apply statistics to the study of geo-ecological processes. Socio-economic statistics are collected and published based on administrative division. So they have low spatial

* Corresponding author.
E-mail address: wangzheng@casipm.ac.cn (Z. Wang).

https://doi.org/10.1016/j.compag.2018.03.031
Received 28 January 2018; Received in revised form 21 March 2018; Accepted 24 March 2018
0168-1699/

resolution and lack of the description to spatial distribution characteristics of socioeconomic statistics. It is difficult to use them for comprehensive analysis of socio-economic data and other data in practical application, which limits their application to geographical research to a great extent. There are three major problems. First, the contradiction between the spatial heterogeneity of geographical elements and the homogeneity of statistics in the same administrative division; Second, the disagreement between landscape scale and statistical scale; Third, the statistical indicators in different regions are inconsistent (Liu and Li, 2012). The spatialization of socio-economic statistics can solve the above problems effectively (Liao and Zhang, 2009).

Numerous studies focused on the spatialization of socio-economic statistics, including spatialization of population (Tobler et al.. 1995, 1997; Sutton et al., 2001; Tian , 2005) and gross domestic product (GDP) statistics (Ebener et al., 2005; Doll et al., 2006; Sutton et al., 2007; Elvidge et al., 1997, 2009a, 2009b and Ghosh et al., 2009). With the rapid development of Remote Sensing (RS) and Geographic Information System (GIS) technology, the spatialization of agricultural production data are frequently studied, mainly including spatialization of crop acreage (Qiu et al., 2003; Leff et al., 2004; You and Wood, 2006; You et al., 2009; Monfreda et al., 2008; Khan et al., 2010; Zhang et al., 2013; Jin et al., 2015; Salmon et al., 2015; Liu et al., 2017) and agricultural production inputs (Potter et al., 2010; Sun et al., 2010; Yan and Pan, 2014). However, there are fewer researches on crop yield spatialization. For instance, Shi et al. used the cultivated land data to spatialize maize yield per unit area statistics by multivariable linear regression model, and got a spatial distribution map of maize yield per unit area in Jilin province (Shi et al., 2011). Liu et al. took population density as the dependent variables and crop yield as independent variables to construct a regression model with the support of land use data. The model was then applied to spatialize provincial-level crop yield statistics, resulting in a distribution map of crop yield of China at 1 km by 1 km in 2000 and the precision of crop yield spatialization results were analyzed from provincial scale down to prefectural scale and county scale (Liu and Li, 2012). But few studies explored the influence of variables scales and partitioning schemes on precision of crop yield spatialization.

As one of frequently-used geo-data processing methods, spatialization of attribute data inevitably results in errors during data processing. Spatialization errors can be reduced by correcting initial spatialization results. Many error modifying methods have been used to correct spatialization errors, such as average correction method (Wu et al., 2015), proportional coefficient correction method (Shi et al., 2016), weight coefficient correction method based on the basic idea that different farmland types have the same weight (Liao and Qin, 2014). However, there are few researches about comparing the pros and cons of different error correction methods. So, in this study we will discuss the influence of some new error correction methods on crop output spatialization and compare them with the existing error correction methods to improve spatialization precision.

This study attempts to simulate the spatial distribution of crop yield in China using land use data with the following objectives: (1) exploring the influence of variables scales on precision of crop yield spatialization; (2) detecting the influence of partitioning schemes on precision of crop yield spatialization; and (3) comparing the pros and cons of different error correction methods.

2. Data

2.1. Data sources

Five datasets are used for this study.

1. County-level and prefecture-level crop yield statistics of China in 2010. The data come from Statistical Yearbook of China in 2011.
2. Land use dataset of China in 2010. The data set is provided by Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn).
3. County-level administrative map of China in 2010. It mostly includes vector data of county-level administrative boundary in China and other attribute data, such as names and codes for administrative divisions.
4. Township-level administrative map of Henan province. It was provided by National Science & Technology Infrastructure of China, Data Sharing Infrastructure of Earth System Science-Data Center of Lower Yellow River Regions (http://henu.geodata.cn). It mostly includes vector data of township-level administrative boundary in Henan province and other attribute data, such as names and codes for the administrative divisions.
5. 227 township-level crop yield statistics in 2010. The data come from 17 county-level Statistical Yearbooks in 2011.

2.2. Preprocessing

1. Collecting area of secondary farmland categories from land cover data. We obtained area of secondary farmland categories (paddy field and dry land. Their classification information are listed in Table 1) at the county level and prefectural level.
2. Matching crop yield statistics with area of secondary farmland categories. We matched the two data based on the consistency of county name and administrative code and obtained 2318 county level data records and 349 prefectural level data records.

3. Simulation of spatial distribution of crop yield

3.1. Research method

Crop output is proportional to farmland area, and different farmland types have different influence on crop output, and multivariate linear regression analysis method (MLRAM) is the most-frequently used method to realize spatialization of attribute data. So, we chose MLRAM to spatialize crop output. Its basic formula is as follows:

Supposing one dependent variable y is affected by k independent variables ($x_1, x_2, \ldots, x_k$), and there are n groups of observed values (y_a, $x_{1a}, x_{2a}, \ldots, x_{ka}$), a = 1, 2, …, n. The general equation of MLRAM can be described as:

$$Y_a = \beta_0 + \beta_1 X_{1a} + ,\ldots,+\beta_k X_{ka} + \varepsilon_a \quad (1)$$

Table 1
The cultivated land classification system of land use of China.

First classification	Secondary classification	Code	Definition
Cultivated land	Paddy field	11	The cultivated land, which has water source guarantee and irrigation facilities and can be normally irrigated in general, and which is used to grow rice, lotus root or other aquatic agricultural crops. And the land with a paddy-upland rotation system is included, too.
	Dry land	12	The cultivated land, which doesn't have water source or irrigation facilities and depends on natural precipitation to grow crop. The dry land, which has water source and irrigation facilities and can be normally irrigated in general. The cultivated land which is mainly used to grow vegetables, and the land which is not be used in the process of normal rotation.