**WFL** Publisher
*Science and Technology*

Meri-Rastilantie 3 B, FI-00980
Helsinki, Finland
e-mail: info@world-food.net

*Journal of Food, Agriculture & Environment Vol.10 (3&4): 1350-1354. 2012* *www.world-food.net*

# Non-parametric method for filling in the missing value for cross-sectional dataset: A validation on the per capita GDP data at county level in China

**Xiangzheng Deng [1, 2*], Yin Fang [1, 3], Yingzhi Lin [4] and Yongwei Yuan [5]**

*[1] Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, No 11A, Datun Road, Anwai, Beijing, 100101, China. [2] Centre for Chinese Agricultural Policy, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China. [3] Graduate School of the Chinese Academy of Sciences, Beijing, 100049, China. [4] School of Mathematics and Physics, China University of Geosciences, No. 388, Lumo Road, Wuhan, 430074, China. [5] Faculty of Resources and Environmental Science, Hubei University, Wuchang District, Wuhan, 430062, China.*
**e-mail: dengxz.ccap@igsnrr.ac.cn*

*Received 22 June 2012, accepted 30 September 2012.*

**Abstract**

When dealing with the observation with missing values, we used to get them by means of mathematical interpolation. Compared with the traditional methods for parametric interpolation including linear interpolation, spline interpolation, kriging interpolation, etc., which sometimes export so paradoxical results that there are quite a lot of debates on the reliability of rationale and application, the non-parametric methods are becoming more and more popular to interpolate the missing values for the cross sectional dataset. In this paper, a non-parametric method is introduced and its feasibility of filling in missing values of per capita GDP data at county level for China is illustrated and verified. The results indicate that the non-parametric method produces essentially unbiased estimates by using kernel density function based on a sample drawn from all the observations. So it appears that the actual performance of non-parametric model can be quite helpful to fill in the missing values with a large sample of observation and the non-parametric extrapolation methods tested in this empirical study could be applied in other similar studies.

***Key words:*** Non-parametric method, stepwise multiple regression, interpolation, per capita GDP, China.

## Introduction

Missing values often plague the spatial analysis with the topics of regional science, e.g. the conglomeration of regional economies. When dealing with a fragmentary data set, the analyst needs to choose a strategy, such as ad-hoc methods (e.g. analysis of the complete cases only, available case methods, and use of some indicator variables with means filled in), likelihood-based approaches (e.g. EM algorithm, structural equations or mixed models) and interpolation-based methods, to fill in the missing values. The relative merits and limitation of these approaches have been discussed elsewhere [1]. Generally, the model-based interpolation algorithm which is most widely applied includes parametric method or non-parametric method. Although the estimated results are easy to explain, the parametric method bears a lot of shortcomings, e.g. as for the form of the model function, once fixed, it is fairly difficult to adjust according to attributes of variables to be included in the estimate. Non-parametric methods, however, can always export the dataset with filled values which are predicted according to those observations without missing values with good fitness, even though its algorithm is sometimes uncertain in the form of regression functions. This paper aims at developing an interpolation method according to the multivariate non-parametric algorithm that has two major advantages compared to the parametric models. Firstly, it fills in the missing values without requiring specification of an explicit functional form. Secondly, it generates statistical measures fit for every point, rather than merely for the full sample.

Socio-economic variables have been widely used in the research fields on regional science as essential model parameters, and at the present the more mature geological models of computing are based on spatial data, for which clearing up the socio-economic attribute data and sampling discrete point data to accommodate the spatial scale geological modeling has become the major task. At present, there are a number of regression-based interpolation methods such as linear interpolation spline interpolation, kriging interpolation and so on, but in some cases their interpolating results can't be favored by the professional scholars after a long period of application [2-4]. To this end, for professional analysis of the data, the interpolation model based on non-parametric algorithm has more and more become one of the research hotspots.

This paper aims to develop a non-parametric interpolation method to fill in the missing values of incomplete data. The non-parametric interpolation method can be used to predict the missing value data according to the overall similarities between the samples without missing values and the samples with missing values and fill in the missing values using the predicted values. The "overall similarities" are measured by the similarities of explanatory variables tightly correlated with the variable with missing values. Considering the missing-value puzzledom is more hackneyed when

dealing with cross-sectional data with social-economic variables, an empirical study on per capita Gross Domestic Production (GDP) at county level and indicators to identify regional economic development and people's living standard, was carried out to illustrate the application as well as verify the validity of the non-parametric interpolation method.

## Material and Methods

***Data:*** As one of the basic measures of a country's economic performance, per capita GDP is always defined as the approximation of the value of goods produced per person in the country in a year, which is the most important indicator to identify the level of economic development [5]. We calculated the variable of the county level, per capita GDP, through dividing the county's GDP by the total number of people in the county of 2005, which comes from National Development and Reform Commission of China (Table 1, row 1).

To identify the "total similarities" between the samples without missing values and the samples with missing values, other variables tightly correlated with per capita GDP are needed. Intuitively, climate, terrain, soil property and land use type are spatially associated with the magnitude and structure of per capita GDP. Therefore, we generated a number of measures (intermediate variables) of climate, terrain, soil property and land use type to assist to fill in the missing values of per capita GDP (Table 1).

*Climatic data:* The climatic data for measuring rainfall (measured by mean annual precipitation in millimeters in 2005), temperature (measured by mean air temperature in 0.1°C in 2005), accumulated temperature (measured by accumulated air temperature ($\geq$°C) in 0.1°C in 2005) is from the data center of Chinese Academy of Sciences (CAS) but were initially collected and organized by the Meteorological Observation Bureau of China from more than 600 national meteorological observation stations. We took the observation data from all stations and interpolated them into surface using an approach called thin plate smoothing splines [6]. Considering there might be non-linear relationships between climatic variables and per capita GDP, we created two new variables, mean annual precipitation square and mean annual mean air temperature square, which are the square terms of mean annual precipitation and mean air temperature, respectively.

*Geographic data:* There were six geographic variables involved in this study. Variables elevation and terrain slope, which measure the nature of the terrain of each county, are generated from China's digital elevation model data set derived from the CAS data center [7]. In addition, we created two measures of distance, distance to provincial capital and distance to port city, to describe the location of each county. A region dummy variable, east, which equals 1 when the county lying in the east of China and 0 other region,was also generated. Share of plain area was created to reflect the landform of each county.

*Land use type:* A land use database developed by the CAS with original data from Landsat TM/ETM images of the year 2005 which have a spatial resolution of 30 m by 30 m was involved in this study. A hierarchical classification system of 25 land use types was originally applied to this database [7, 8]. Among all of the land use types, built-up area was aggregated firstly into one km by one km picture elements ('pixels') and then to the county level as variable built-up area, which was the observation used in this study.

*Other socio-economic data:* While investment in agriculture sector is indirectly determined by local government earning which is tightly related to the per capita GDP, and population of each county owns direct relationship with per capita GDP, so two available social-economic variables, investment in agriculture sector and population were introduced in our study. In addition, the variable highway density was created to reflect the transportation of each county.

To prevent the insignificant relationship between intermediate variables and per capita GDP from weakening the robustness of interpolation, we farther chose some explanatory variables from the intermediate variables statistically significant related with per capita GDP for our analysis by using a strategy of stepwise regression. The selected explanatory variables were introduced to calculate the weight and used to fill in the missing value of per capita GDP.

***Methods:*** The interpolation method specified in this section is practically a non-parametric estimation model. Specifically, the kernel density function is chosen and employed to build the non-parametric interpolation model which is viewed as the calculation of a weighted sample mean [9, 10]. In other words, the missing values $Y$ of pending points can be estimated using the following equation:

$$Y = \sum_{i=1}^{n} W_i(X; X_1, X_2, \ldots, X_n) Y_i \quad (1)$$

where $\{Y\} \cup \{Y_i\}$ consists of the data series of the cross sectional dataset, $Y$ represents the missing values to be filled in while $Y_i$ is observation values; $X$ and $X_i$ are vectors constructed by explanatory variables of the pending and observed points, respectively; $W_i(X; X_1, X_2, \ldots, X_n)$ is the weight function which meets $W_i(X; X_1, X_2, \ldots, X_n) \geq 0$ and $\sum_{i=1}^{n} W_i(X; X_1, \cdots, X_n) = 1$.

The weight function $W_i(X; X_1, X_2, X_n)$ is calculated by kernel function $K$ (Equation 2):

**Table 1.** Descriptive statistics of key variables at county levels used in this study.

| Variables | Units | Mean | Std. |
|---|---|---|---|
| Per capita GDP | yuan | 4957.13 | 8622.26 |
| Mean air temperature | 0.1°C | 12.68 | 5.22 |
| Accumulated air temperature(≥0℃) | 0.1°C | 4960.46 | 1429.14 |
| Mean annual precipitation | 0.1mm | 991.03 | 450.72 |
| Elevation | meter | 600.04 | 713.82 |
| Terrain slope | degree | 2.75 | 2.81 |
| Share of plain area | 0.01% | 0.35 | 0.38 |
| Distance to provincial capital | km | 171.56 | 109.67 |
| Distance to port cities | km | 507.44 | 346.86 |
| East | - | 4960.46 | 1429.14 |
| Built-up area | hectare | 8.82 | 0.99 |
| Investment in agricultural sector | yuan per capita | 9.77 | 55.19 |
| Population | persons per county | 569629.10 | 518959.70 |
| Highway density | meter/ hectare | 34.74 | 103.09 |

$$W_i(X;X_1,\cdots,X_n)=K\left(\frac{X-X_i}{a_n}\right)\Big/\sum_{i=1}^{n}K\left(\frac{X-X_i}{a_n}\right) \quad (2)$$

Here, the kernel function is defined as the density function of normal distribution (Equation 3):

$$K(\mu_i)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(\mu_i)^2),\quad \mu_i=\frac{X-X_i}{a_n} \quad (3)$$

where $a_n$ is the bandwidth. At this point, Equation 1 could be written as follow:

$$Y=\sum_{i=1}^{n}\left[K\left(\frac{X-X_i}{a_n}\right)\Big/\sum_{j=1}^{n}K\left(\frac{X-X_i}{a_n}\right)\right]Y_i \quad (4)$$

The estimated point lays a "soft blanket" on the observations so that it absorbs the peaks of the highest poles (upward outliers) and remains above the lowest ones. When emphasis is given to nearby points, the kernel function is said to use a small bandwidth. It is possible to control the bandwidth in order to meet a certain optimality criterion. The larger the bandwidth is, the tighter and smoother the blanket is. Thus the non-parameter interpolation methods allows for a manipulation of the observation errors by the degree to which the surface is smoothed. The bandwidth can be scaled by the user relative to the optimal bandwidth defined as follows:

$$a_n=\left[\frac{4}{n(d+2)}\right]^{\frac{1}{d+4}} \quad (5)$$

with $n$ being the number of observations and $d$ being the number of explanatory vari-ables. In this paper, $Y$ is the per capita GDP at country level and those variables intuitively related with per capita GDP are included the vector $X$.

## Results

***Creating the samples:*** An unabridged dataset with 2063 samples was originally introduced (Table 1, row 1), from which a validation subset was picked up to evaluate the accuracy of non-parametric interpolation method. Concretely, 601 of the 2063 counties of China were randomly chosen (Fig. 1 panel a), for which the variable per capita GDP was estimated by non-parametric interpolation method and compared with their actually observed values. Other 1462 samples are the observations used to estimate and fill in the 601 selected values of per capita GDP in this study.

***Choosing the explanatory variables:*** We firstly carried out the stepwise multiple regression analysis between per capita GDP (dependent variable) and intermediate variables (independent variables) to pick up variables significantly correlated with per capita GDP. The estimation results show that three of the climatic variables, mean annual precipitation, mean annual precipitation square and mean air temperature square, are significantly correlated with per capita GDP (Table 2, row1, 2 and 3). As for geographic variables, elevation, share of plain area and east are all positively related with per capita GPD at significant level of 0.01 (Table 2, row 4, 5 and 6). Other intermediate variables, including built-up area which belongs to land use type variables, investment in agricultural sector, population and highway density which belong to social-economic variables, all maintain significantly positive relationships with per capita GDP (Table 2, row 7, 8, 9 and 10).

**Table 2.** Estimates of stepwise multiple regression.

| | Dependent variable: per capita GDP |
|---|---|
| Mean annual precipitation | 2.62 (4.68)*** |
| Mean annual precipitation square | -0.003 (2.23)*** |
| Mean air temperature square | 25.794 (1.80)* |
| Elevation | 1.45 (3.57)*** |
| Share of plain area | 2540.56 (3.22)*** |
| East | 3989.42 (8.15)*** |
| Built-up area | 2726.49 (11.33)*** |
| Investment in agricultural sector | 17.41 (4.88)*** |
| Population | 0.01 (3.57)*** |
| Highway density | 15.16 (7.55)*** |
| Constant | -16458.60 (-8.97)*** |
| Observations | 1462 |

***Interpolating result and validation:*** The per capita GDP of 601 counties selected are filled in based on the other 1462 observations and the relationship between per capital GDP and the explanatory variables by adopting the non-parametric interpolation method. It shows that the spatial pattern of observations with interpolated per capita GDP (Fig. 1 panel b) is kept almost the same as that of the 1462 observations (Fig. 1 panel b).

After drawing the probability density histogram, we find that the per capita GDP data before and after interpolation both meet the normal distribution (Fig. 2). Both of the probability density histograms almost take the same forms (Fig. 2 panel a versus panel b) which shows that the filling-in process does not apparently change the distribution of the per capita GDP.

To argue that the interpolation result is robust and accurate, a comparative analysis was implemented between the values filled in and the values originally observed of the variable per capita GDP. A statistic of standard difference of the values filled in and the values originally observed of the variable per capita GDP was created, for which three-sigma principle is applicable (Fig. 3). The graph of the standard difference shows that all the points are located in the range between 3 (upper critical line) and -3 (lower critical line), which indicates that there do not exist abnormal points. In this sense, we conclude that the interpolation accuracy is reliable.

## Discussion and Conclusions

Methodologically, this paper has introduced a non-parametric interpolation method and illustrated how it could be used to fill in the missing values for the cross section dataset. Compared with the parametric methods, the non-parametric approach which makes

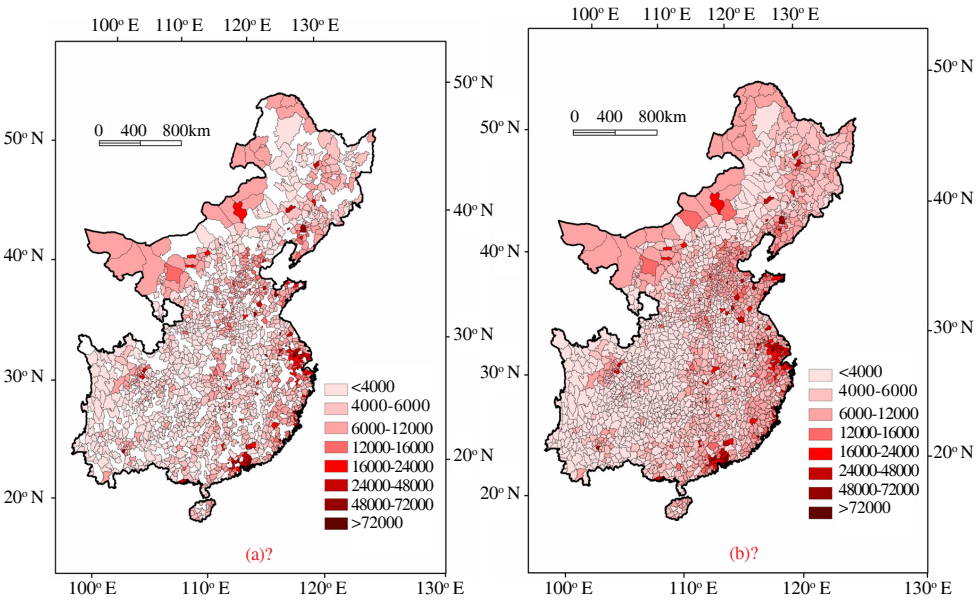


***Figure 1.*** Map of per capita GDP at county level in China: (a) with missing values; (b) with interpolated values.

Note: After looking in as many published sources as possible (both national and provincial yearbooks and statistical compendia), the measurement of GDP were not ready to use for a large number of counties (or county level administration) in Taiwan, Hong Kong, Macao, Xinjiang, Tibet, Qinghai and Gansu. Therefore, in this study, we eliminate the counties from Xinjiang, Tibet, Qinghai and Gansu from the analysis we are left with a final sample size that includes 2063 counties.

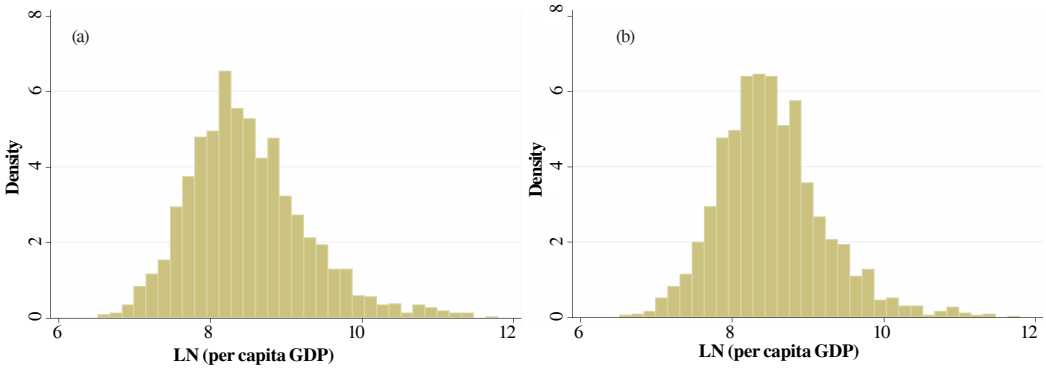


***Figure 2.*** Probability density distribution by a kernel density estimator for the measurement of per capita GDP at counties of China: (a) with missing values; (b) with interpolated values.

Note: All China's counties excluding Taiwan, Hong Kong, Macao, Xinjiang, Tibet, Qinghai and Gansu are included (including prefectural and provincial capitals). For simplicity, we herein still call them "counties." All numbers in 2005 real terms.

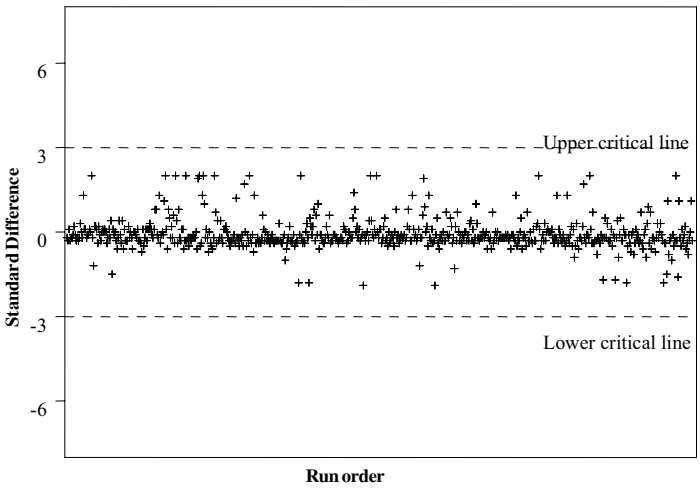


***Figure 3.*** Standard difference graph for validating the interpolation.

Note: Standard Difference = [$D$ –mean($D$)]/std.($D$), where $D$ is the difference of the values filled in and the values originally observed of the variable per capita GDP; mean($D$) is the mean of $D$; and std.($D$) is the standard deviation of $D$.

the non-linearity forms of functions embedded in the interpolation methods, owns the advantage that it gives the measure of statistical reliability at every point. In addition, the reliability of the estimated values can be easily diagnosed and evaluated.

Given the advantages of non-parametric interpolation method, we used it to fill in the missing values of per capital GDP at county level in China. By filling the values of per capita GDP of a selected validation subset, we illustrated the application of the non-parametric interpolation method. The interpolation results show that this approach keeps the spatial pattern as well as the probability distribution of interpolated variables. The validation for the interpolation results in this study also demonstrates that the interpolation is robust and accurate.

Although have been improved reliable, the non-parametric interpolation method is criticized due to the limitation that it is "weak on theory" in that the resulting regression curve is shaped according to the data but the imposed properties of theoretical functions. This might be unacceptable to some modelers and experts. However, the idiosyncrasy of not requiring specification

of an explicit functional form endows the desirable with the virtue that the non-parametric method does not depend on the fit at other points.

**Acknowledgements**

This research was supported by the National Basic Research Program of China (2012CB95570001; 2010CB950904); Data supports from projects funded by the National Science Foundation of China (40801231; 41071343; 41171434; 70873118) and National Soft Science Research Program of China (2010GXS5B163) are also appreciated.

**References**

[1]Little, R. J. A. and Rubin, D. B. 2002. Statistical Analysis with Missing Data. 2nd ed. Wiley,New York.

[2]Johnson, C. R. and Smith, R. L. 2001. Linear interpolation problems for matrix classes and a transformational characterization of M-matrices. Linear Algebra and its Applications **330**(1-3):43-48.

[3]Nürnberger, G. and Zeilfelder, F. 2000. Developments in bivariate spline interpolation. Journal of Computational and Applied Mathematics **121**(1-2):125-152.

[4]Gatti, M. N., Milocco, R. H. and Giaveno, A. 2003. Modeling the bacterial oxidation of ferrous iron with Acidithiobacillus ferrooxidans using kriging interpolation. Hydrometallurgy **71**(1-2):89-96.

[5]Mankiw, N. G. 2002. Macroeconomics. Worth Publishers,New York.

[6]Hartkamp, A. D., Kirsten, D. B. and Stein, A. 1999. Interpolation Techniques for Climate Variables. CIMMYT, Mexico D.F.

[7]Deng, X. Z., Huang, J. K., Rozelle, S. and Uchida, E. 2008. Growth, population and industrialization, and urban land expansion of China. Journal of Urban Economics **63**(1):96-115.

[8]Beirens, H. J. 1987. Kernel Estimations of Regression Functions. Advances in Economics 6, Cambridge University Press.

[9]Keyzer, M. A. and Sonneveld, B. G. J. S. 1997. Using the mollifier method to characterize datasets and models: the case of the Universal Soil Loss Equation. ITC Journal **3-4**:263-272.

[10]Rubin, D. B. 1996. Multiple imputation after 18+ years. Journal of the American Statistical Association **91**:473-489.