RESEARCH ON GPS HEIGHT FITTING BASED ON LINEAR REGRESSION MODEL

Kunyu Yang 1,2, Lilong Liu 1,2, Liangke Huang 1,2,3,4

1 College of Geomatic Engineering and Geoinformatics, Guilin University of Technology, Guilin, 541004, China.
2 Guangxi Key Laboratory of Spatial Information and Geomatics, Guilin, 541004, China.
3 GNSS Research Center, Wuhan University, Wuhan, 430079, China.
4 School of Geodesy and Geomatics, Wuhan University, Wuhan, 430079, China. (email: lkhuang@whu.edu.cn) or L.H. (email: ybyao@whu.edu.cn)

KEYWORDS: GPS elevation fitting; linear regression; polynomial fitting model; fitting accuracy;

ABSTRACT:

This paper mainly expounds the parameter estimation method, the outlier diagnosis and the establishment of the optimal regression equation in the linear regression model theory, the analysis of the principle of the polynomial fitting model, the derivation of the algorithm process, and the research on the accuracy evaluation method. The GPS survey area is fitted and calculated. The fitting model is analyzed and compared in detail. The better parameter values and regression equation models of the planar region are estimated. The fitting accuracy meets the requirements of the fourth level measurement, which can be used in actual engineering. Replace the fourth level measurement in the application.

1 Introduction

Using the general linear hypothesis theory in the linear regression model, the significance of regression equations and regression coefficients in linear models, especially the gross error detection theory, is established, and a robust linear regression model is established to improve the prediction accuracy. Because the purpose of modeling is to fit the elevation anomalies that have not been leveled, a reliable geoidal contour map of the study area is established. Fully understand the selection criteria of the evaluation regression equation, and according to different needs, according to different optimal equations, the optimal regression equation is established. Emphasis is placed on the use of stepwise regression theory to discuss different linear model building methods to improve the applicability of the model.

2 Regression model overview

Linear statistical model is a kind of highly practical model. It is also widely used in the processing of surveying data. Especially in the process of elevation fitting, many methods use it, such as moving surface method and polynomial fitting method. The theory mainly includes important parts such as parameter estimation, hypothesis analysis, linear regression, etc. The following mainly discusses three aspects:

2.1 Research on theoretical modeling of parameter estimation

The study uses the theory of parameter estimation in linear regression model theory, such as regression diagnosis (Cook distance) and Box-Cox transformation to improve the accuracy and reliability of parameter estimation in linear transformation model. The estimation of the regression parameters, the parameter estimation mainly solves the regression diagnosis statistic by knowing the point data, and analyzes the abnormal points by comparing the sizes of the respective quantities, thereby ensuring the accuracy requirement of the elevation fitting.

2.2 Hypothesis testing

2.2.1 Significance test of regression equation

The so-called significance test of the regression equation is to test the hypothesis: all regression coefficients are equal to zero, which is the test

\[ H_0 : \beta_1 = \ldots = \beta_p = 0 \]  

\[ H_1 : \beta_1 \neq 0 \cup \beta_2 \neq 0 \cup \ldots \cup \beta_p \neq 0 \]  

If we conclude that we reject the null hypothesis H0, this means that we accept H1: at least one \( \beta_i \neq 0 \). On the contrary, if the conclusion of the test is to accept the null hypothesis H0, this means all \( \beta_i = 0 \), that is, for the error, The effect of any independent variable on the dependent variable is not important. The construction test statistic is:

\[ F = \frac{SS/(p-1)}{RSS/(n-p)} \]  

Among them \( \beta_i = 0 \) , when the null hypothesis (1) is established, \( F \sim F_{p-1,n-p} \) , for a given level \( \alpha \) ,

1 Corresponding author: Lilong Liu; E-mail: ln_liulilong@163.com
when \( F > F_{p-1,n-p}(\alpha) \), we accept the alternative hypothesis \( H_1 \), otherwise we accept the null hypothesis \( H_0 \).

### 2.2.2 Significance test of regression coefficient

The significance test of the regression equation is a holistic test of linear regression. If we test the result of rejecting the null hypothesis, this means that the dependent variable \( Y \) depends linearly on the independent variable \( X_1, \ldots, X_p-1 \), which is the regression independent variable. Overall, but this does not exclude that \( Y \) does not depend on some of these arguments, some \( \beta_i \) could be equal to zero. Therefore, when the significance test of the equation is rejected, we also need to make a significant hypothesis test for each independent variable one by one, that is, the fixed \( i, 1 \leq i \leq p-1 \) is tested as follows:

\[
H_0^i: \beta_i = 0, \quad H_1: \beta_i \neq 0 \tag{4}
\]

For models

\[
y = \mathbf{x} \beta + \mathbf{e}, \quad \mathbf{e} \sim N(0, \sigma^2 \mathbf{I}) \tag{5}
\]

The least squares estimate about \( \beta \) is \( \hat{\beta} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} \), if \( c_{ij} = (\mathbf{X}' \mathbf{X})^{-1} \), according to the theorem:

\[
\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii}) \tag{6}
\]

So when \( H_0^i \) is established

\[
\frac{\hat{\beta}_i}{\sqrt{c_{ii}}} \sim N(0,1)
\]

Because of

\[
\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p}, \quad \text{and it is independent with} \quad \hat{\beta}_i
\]

According to the definition of the \( t \) distribution:

\[
t_i = \frac{\hat{\beta}_i}{\sqrt{c_{ii}}} - t_{n-p} \tag{7}
\]

For a \( \alpha \) that have been given level, at the time

\[
t_i > t_{n-p}(\alpha/2)
\]

we accept the alternative hypothesis \( H_1 \); otherwise, we accept the null hypothesis \( H_0 \).

### 2.2.3 Gross error detection theory

There are two general ideas for the method of gross error detection. One is to use the gross error as the parameter to be estimated, the idea of quasi-stationary adjustment is adopted to solve the rank-deficient problem and then directly obtain the gross error, and the other is to select some observations as the standard. For observations, it is proposed to calculate the parameters to be estimated by the least squares method, and treat the residuals of the non-quasi-observed values as gross errors.

Assume that the error equation is:

\[
\mathbf{e} = \mathbf{Y} - \mathbf{X} \hat{\beta}
\]

The original hypothesis of the data detection method is

\[
H_0: E(e_i) = 0, \quad \text{means that there is no gross error in the observed values} \quad x_i,
\]

so consider

\[
e_i \sim N(0, \sigma^2 e_i^2)
\]

It can be used as a standard normal distribution statistic:

\[
u = \frac{e_i}{\sigma e_i}
\]

For \( u \) test, if

\[
u > u_{\alpha/2}
\]

then negation \( H_0 \),

\[
E(e_i) = 0, \quad \text{there may be gross errors in} \quad x_i.
\]

### 3. Optimal regression equation design

For the regression equation, the so-called selection regression equation mainly consists of two parts. The first point is the choice of regression equation types, that is, their relationship is linear and nonlinear when solving specific problems. The second point is what the independent variables choose after the model has been determined. When we determine that the dependent variable and the independent variable that may affect it are suitable for a linear regression model, the result is that all independent variables, some even independent variables that have no effect on the dependent variable, are included in the regression equation, resulting in the amount of calculation becomes large, and the accuracy of the forecast also drops a lot. Therefore, when applying regression analysis to solve practical problems, it is very important to select an optimal subset of independent variables from the set of independent variables that maintain a linear relationship with the dependent variable.

Steps to specifically select the optimal regression equation:

1. The regression variable set \( X \) \((1-n)\) is divided into two parts, one is an important set of independent variables, denoted as \( X_1 = \{x_1, x_2, \ldots, x_k\} \), and the other part is a set of variables to be selected, denoted as \( X_2 = \{x_{k+1}, x_{k+2}, \ldots, x_n\} \). (2) Specially write \( m=1,2, \ldots, k \) independent variables in \( X_1 \) into the equation, use stepwise regression to filter the remaining variables in the remaining \( n-m \), and control the total number of variables written into the equation. Within seven.

2. For example, if the number of important independent variables is \( k=4 \), then the equations that can be established are \( 24-1=15 \).

3. Perform statistical tests on the above equations, test the test and the live test, and then combine the results of the test to select the optimal regression equation.

As can be seen from the foregoing, the magnitude of the
residual squared RSS reflects the degree of deviation between the actual data and the theoretical model. It is an important criterion for evaluating the regression equation. Generally speaking, the smaller the RSS, the better the data and the model fit, assuming the full model is:

\[ Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{q-1} X_{q-1} + e \]  

(10)

Then the sum of squared residuals is:

\[ RSS_q = y'(I - X_q (X_q' X_q)^{-1} X_q') y \]  

(11)

Since RSSq decreases as q increases, in order to prevent too many independent variables to be selected, we multiply the sum of squared residuals by a function that increases with a q as a penalty factor, which is recorded as:

\[ RMS_q = \frac{1}{n-q} RSS_q \]  

(12)

According to the nature of RMSq, we can select the subset of independent variables according to the principle that the RMSq is smaller, and it is called RMSq criterion.

The same similar guidelines are the Cp guidelines proposed by Mallows in 1964, which are:

\[ C_p = \frac{RSS_q}{\sigma^2} - (n-2q) \]  

(13)

And the AIC guidelines proposed by Japanese statistician Akaike in 1974, namely:

\[ AIC = n \ln(RSS_q) + 2q \]  

(14)

The selection principle of the three criteria is as small as possible. The optimal subset is selected by assuming the elevation anomaly and the full model between the X and Y independent variables, and then the optimal regression model is constructed to fit the elevation of the point to be solved.

4 survey area data verification

4.1 Source of experimental data

The experimental data is derived from the GPS control network in a gentle riverside area. The control network has a rating of B and a normal elevation is measured at a second level. The specific values are shown in the table below:

| serial number | X      | Y      | H      | Hr     | \( \zeta \) |
|---------------|--------|--------|--------|--------|-----------|
| 2             | 3565858.080 | 499248.000 | 31.225 | 10.181 | 8         |
| 3             | 3564029.592 | 499613.378 | 30.134 | 9.0686 | 21.0654   |
| 4             | 3565549.066 | 498813.558 | 31.110 | 10.081 | 5         |
| 5             | 3563826.318 | 499348.917 | 30.059 | 9.0030 | 21.0606   |
| 6             | 3566091.401 | 499632.434 | 30.709 | 9.6536 | 21.0554   |
| 7             | 3564312.203 | 500321.498 | 31.206 | 10.120 | 4         |
| 8             | 3566375.346 | 499179.740 | 26.424 | 5.3896 | 21.0856   |
| 9             | 3564827.161 | 500392.773 | 29.487 | 8.4044 | 21.0826   |
| 10            | 3566324.251 | 498659.474 | 25.790 | 4.7728 | 21.0821   |
| 11            | 3564001.762 | 500035.270 | 31.338 | 10.2591| 21.0789   |
| 12            | 3567660.247 | 499189.334 | 31.128 | 10.1110| 21.0701   |
| 13            | 3566814.699 | 499080.199 | 30.989 | 9.9617 | 21.0273   |
| 14            | 3567961.396 | 498691.018 | 30.160 | 9.1622 | 20.9978   |
| 15            | 3566854.849 | 498567.506 | 30.741 | 9.7334 | 21.0076   |
| 16            | 3567538.805 | 499757.287 | 31.074 | 10.010 | 1         |
| 17            | 3566774.076 | 499599.745 | 30.621 | 9.5500 | 21.0710   |
| 18            | 3567524.909 | 500219.017 | 31.183 | 10.131 | 4         |

Table 1 Raw data of elevation anomalies
4.2 Height fitting data calculation

There are 18 points in the survey area, and 7 points are selected as the fitting points, which are 5, 9, 11, 2, 15, 16, and 12 respectively. The remaining points are used as check points. Use $C_a$ to indicate its distribution as:

![Figure 1 Distribution of GPS points](image)

### 4.2.1 Calculation of residuals and cook statistics

| serial number | $e_i$  | $r_i$  | $h_i$  | $D_i$  |
|---------------|--------|--------|--------|--------|
| 1             | -0.0009| -0.1110| 0.1472 | 0.0007 |
| 2             | 0.0028 | 0.3155 | 0.0664 | 0.0024 |
| 3             | 0.0004 | 0.0517 | 0.1689 | 0.0002 |
| 4             | 0.0017 | 0.2062 | 0.1685 | 0.0029 |
| 5             | -0.0008| -0.0961| 0.2357 | 0.0002 |
| 6             | 0.0027 | 0.3017 | 0.0652 | 0.0021 |
| 7             | -0.0030| -0.3600| 0.2039 | 0.0116 |
| 8             | -0.0003| -0.0393| 0.0703 | 0.0000 |
| 9             | -0.0053| -0.6573| 0.2084 | 0.0379 |
| 10            | 0.0008 | 0.0951 | 0.1746 | 0.0006 |
| 11            | -0.0014| -0.1659| 0.1770 | 0.0020 |
| 12            | -0.0100| -1.1957| 0.1460 | 0.0815 |
| 13            | -0.0011| -0.1298| 0.0900 | 0.0006 |
| 14            | -0.0095| -1.1795| 0.2154 | 0.1273 |
| 15            | -0.0022| -0.2694| 0.1952 | 0.0059 |
| 16            | 0.0157 | 1.9432 | 0.2065 | 0.3275 |
| 17            | 0.0237 | 2.7405 | 0.0954 | 0.2639 |
| 18            | -0.0132| -1.8228| 0.3653 | 0.6376 |

Table 2 Diagnostic statistics of elevation fitting data

As can be seen from the above table, the residual value of point 17 $e_{17} = 0.0237$, $r_{17} = 2.7405$, which is significantly larger than the absolute value of the corresponding amount of other points, which indicates that the 17th data will be far away from other points on the residual map, showing an abnormality. However, in the cook statistic, $D_{17} = 0.2639$ is much smaller than $D_{18} = 0.6376$, and other values $D_i$ and $D_{18}$ are also very small. Therefore, the 18th data is a data that has a great influence on the regression estimation, and needs special attention. After carefully comparing the data, I found that the transcript was correct, so the point should be removed.

### 4.2.2 Box-Cox transformation calculation

1. Before the data is normalized, we can use the box-cox transformation to get the following table:

| lambda | -2   | -1   | -0.5 | 0    | 0.125 | 0.25 |
|--------|------|------|------|------|-------|------|
| $RSS$  | 0.048491 | 0.000404 | 0.000056 | 0.000014 | 0.000011 | 0.000009 |
| $RSS$  | 0.000008 | 0.000007 | 0.000007 | 0.000006 | 0.000006 | 0.000005 |

Table 3 Correspondence between transformation parameters and $RSS$

From the above table, we can see that when lambda=2, the residual squared sum $RSS$ is the smallest, so we can approximate that lambda=2 is the optimal choice of transform parameters. At this time, the regression model selected according to this parameter can find the following results:
The internal accuracy and the external accuracy are equal, $u=m=1.0335$ (the fitting accuracy is too bad, invalid)

2. After the data is normalized, we can get the following table:

| lambda | -2   | -1   | -0.5 | 0    | 0.125 | 0.25 |
|--------|------|------|------|------|-------|------|
| $RSS$  | 0.007035 | 0.002432 | 0.001810 | 0.001114 | 0.000928 | 0.000743 |
| $RSS$  | 0.000562 | 0.000392 | 0.000241 | 0.000011 | 0.000020 | 0.001944 |

Table 4 Correspondence between transformation parameters and $RSS$

From the above table, we can see that when lambda=0.75, the residual square sum $RSS$ is the smallest, so we can approximate that lambda=0.75 is the optimal choice of transform parameters. The correspondence diagram is:
Among them, the internal accuracy is $u=0.0048$, the external conformity is $m=0.0048$, and the internal precision $q=0.0008998$.  

### 4.2.3 Significance test of regression equation

According to the method described above, we can draw the following table:

| Source of variance | Quadratic sum $u$ | Degree of freedom $m$ | Mean square $q$ | F-ratio |
|--------------------|------------------|----------------------|-----------------|--------|
| Regression         | 0.0022           | 2                    | 0.0011          | 897.4346 |
| Error              | 0.000004858      | 4                    | 0.00001215      |        |
| Aggregate          | 0.002204858      | 6                    |                 |        |

Table 6: Analysis of variance

Given a level $\alpha = 0.005$, $F_{2,4}(0.005)=26.3$, and $F>F_{2,4}(0.005)$, we reject the null hypothesis that the elevation anomaly has a certain dependence on the X, Y coordinates.

### 4. Choice of optimal regression equation

Let the full model of the polynomial fit be:

$$\zeta = a_1 + a_2 x + a_3 y + a_4 x^2 + a_5 y^2 + a_6 xy$$

Where $x$, $y$ are indispensable variables, then there are 7 subsets of variables, all the results are shown below:

| Variable subset | RMSq | Cp   | AIC   |
|-----------------|------|------|-------|
| $x, y$          | 0.00000121 | 9.9743 | -79.6445 |
| $x, y, x^2$     | 0.00000106 | 8.1965 | -80.5982 |
| $x, y, y^2$     | 0.00000026 | 2.7537 | -90.4813 |
| $x, y, x^2, y^2$| 0.00000022 | 4.0997 | -92.3458 |
| $x, y, x^2, xy$ | 0.00000141 | 9.3900 | -79.4302 |
| $x, y, y^2, xy$ | 0.00000025 | 4.1157 | -91.6472 |
| $x, y, x^2, y^2, xy$ | 0.00000044 | 6.0000 | -90.4134 |

Table 7: RMSq, Cp, and AIC values for all possible regressions of the fitted data.
According to the above table, when the subset of variables is \([x, y, x^2, y^2]\), the values of RM Sq and AIC criteria are the smallest, so the regression model established by this subset is the optimal regression model. The results after fitting according to the model are:

| X  | Y  | H  | Hr | z  | z* | v  |
|----|----|----|----|----|----|----|
| 356400.762 | 5080035.270 | 31.338 | 10.2591 | 21.0789 | 21.0671 | 0.0118 |
| 3567660.247 | 4998189.334 | 31.128 | 10.1110 | 21.0170 | 21.0263 | -0.0093 |
| 3566814.699 | 499080.199 | 30.989 | 9.9617 | 21.0273 | 21.0282 | -0.0009 |
| 3566854.849 | 498567.506 | 30.741 | 9.7334 | 21.0076 | 21.0083 | -0.0007 |
| 3566735.346 | 499179.740 | 26.424 | 5.3896 | 21.0344 | 21.0342 | 0.0002 |
| 3568427.161 | 5080392.773 | 29.487 | 8.4044 | 21.0826 | 21.0755 | 0.0071 |
| 3566324.251 | 498659.474 | 25.790 | 4.7728 | 21.0172 | 21.0149 | 0.0023 |

Table 8 Elevation anomalies and residuals after fitting

Among them, the internal accuracy is \(u=0.0069\), the external conformity is \(m=0.0069\), and the internal precision \(q=0.0002729\). When the checkpoint is fitted according to the full model, the accuracy is \(m=0.0078\).

When we perform the box-cox transformation of the elevation fitting dependent variable, the centralization of the data sample easily leads to the final parameter selection of 1, which is the plane fitting. At this time, you need to change the data processing method, that is, normalize the data. There is no direct and inevitable relationship between the normalization of the data and the selection of the box-cox transformation parameters. Different transformation parameters mean different fitting models, and the final conversion accuracy is different. In general, the accuracy of normalizing the data will be higher. For example, in the Box-Cox transformation, the optimal parameter selection after data centering is 1, and the fitting result according to this parameter is \(u = 0.0192, m = 0.0109\), and the optimal parameter after normalization is 2. The fitting result is \(u = 0.0142\) and \(m = 0.0081\). Obviously, the accuracy of the fitting after the data is normalized is better.

**CONCLUSIONS**

In this paper, different GPS elevation fitting models based on linear regression theory are used to optimize the geoid-like surface and achieve high-precision conversion from high ground to normal high. Through the diagnostic statistic in the regression diagnosis, including the residual and cook statistic, the abnormal data in the control point can be accurately found, and the transcript is checked and eliminated, thereby reducing the influence on the fitting. Box-Cox is a parameter transformation from a comprehensive perspective, which makes the error obey the normal distribution. By selecting the optimal parameter value by comparing the magnitude of the residual value, the complex nonlinear elevation anomaly problem is transformed into Linear relationship to deal with. This paper proves the feasibility and reliability of linear regression model in GPS elevation fitting with concrete examples, which can better reflect the trend and regularity, achieve effective GPS elevation fitting, and meet the accuracy requirements of general measurement in practical engineering. This has a good reference for the wide application of GPS elevation in practical engineering. However, the polynomial surface fitting model based on linear regression theory discussed in this paper mainly refers to the fitting application in a small range. For the measurement area with larger area and more complicated terrain conditions, the application and fitting effect of the model need further discussion.

**ACKNOWLEDGEMENTS**

(This work was sponsored by the National Natural Foundation of China (41664002;41704027); Guangxi Natural Science Foundation of China(2018GXNSFAA294045; 2017GXNSFDA198016; 2017GXNSFA198139);the “Ba Gui Scholars” program of the provincial government of Guangxi; and the Guangxi Key Laboratory of Spatial Information and Geomatics (14-045-24-10; 16-380-25-01))

**REFERENCES**

[1] Chen Shaohua. Research on SINS/GNSS/CNS autonomous navigation technology for high-altitude mobile spacecraft [D]. Nanjing University of Aeronautics and Astronautics, 2012.

[2] Jian Chenghang, Lei Kunchao, Ma Jing, Lin Pei.Application of GPS Height Fitting Method in Engineering[J].Urban Geology,2014,9(01):50-53.

[3] Wang Wei. Research on refinement of mining area like geoid based on regional ellipsoid method [D]. Chang'an University, 2014.

[4] Guo Yang. Feasibility study of GPS elevation fitting to replace the third and fourth leveling measurements [J]. Mine Survey, 2018, 46 (06): 110-112+116.

[5] Jia Xue, Xu Wei, Liu Chao, Zhao Xingwang, Yu Xuexiang. GNSS Elevation Conversion Method Considering Earth's Gravity Field Model[J].Science of Surveying and Mapping, 2019,44(05):14-20.
[6] Li Jian. Research on regional GPS elevation anomaly fitting and modeling method [D]. Kunming University of Science and Technology, 2013.

[7] Qiyake. Application of GPS elevation conversion in highway elevation control measurement [J]. Inner Mongolia Coal Economy, 2015(11): 8-9.

[8] Guan Zhen, Shi Fengqi, Tang Xiurong. Application of Multi-faceted Function Method in GPS Leveling Fitting in Hilly Areas [J]. Surveying and Spatial Geography Information, 2012, 35(01): 124-126.

[9] He Qingqing, Shi Changwei. GPS Height Conversion Model and Accuracy Assessment [J]. Prospecting Science and Technology, 2018(05): 27-30.