Statistical Data Combined Forecasting Model Based on Optimal Estimation of Error Value

Yafen Wang
Zhejiang Technical Institute of Economics, Hangzhou Zhejiang, 310018, China

Abstract
The accurate and reliable statistics are the foundation to grasp the operation of the economy and make scientific decisions. This paper is on the basis of the characteristics of the change of the time series data itself, combined with the characteristics of the grey forecasting model, regression combined model and the double exponential smoothing model, build the combined forecasting model based on the minimum error absolute value sum. Prediction is made on GDP data by the application of the model, and the prediction value obtained represents the "True value", and then from the angle of the abnormal value to analyze the accuracy of the GDP data of China, and the results show that the combined forecasting model is of relatively high practical value in the accuracy test of statistical data, and is worth further study.

Keywords: Statistical Data Quality, Abnormal Value, Combined Forecasting Model.

1. INTRODUCTION
In recent years, the reliability of the statistics data released by Chinese authorities has always been a hot issue which numerous domestic and international research institutions and scholars have debated about. At present, there have been many scholars at home and abroad who adopt different methods to make diagnosis on the reliability of the Chinese statistics. The logic evaluation based on correlation and the evaluation based on anomalous value is the most used method. Wherein, the logic evaluation method based on correlation is in the indicators with high correlation to the statistics data, to make evaluation comment from the known correct indicators on the reliability of the statistics. For example, the principal component regression method has selected 15 indicators from independent sources, with as much correlation with the economic growth as possible, to indicate that the correlation relation between the change of the 15 indicators and the GDP growth is completely in line with economic laws, thereby, it is considered that the economic growth data is reliable (Baa, Besjes, Cote, Koutsma, Lorenz and Short, 2015; Bianchi, Gagliardi, Campanella, Landi, Capaldo, Carleo, Armini, 2013). The evaluation method based on anomalous value is to make use of the statistical method to identify the anomalous values, and combine with the background that generates the anomalous values to judge whether it is created by the statistical data quality problems. For example, from the classic least square method (OLS) estimate the production function can be obtained, and through the residual error, D statistics of Cook, DIFITS statistics and other classical diagnostic statistics, the diagnosis on the abnormal points that are existed in the GDP data have been carried out (Nguyen, Cressie and Braverman, 2012; Rougier, 2013).

The evaluation methods adopted by the aforementioned research are all based on the traditional average, covariance and correlation coefficient matrix, as well as the ordinary least squares method (OLS). However, these traditional statistical methods are not very solid, which are easily affected by a few anomalous values in the data set (Jinnagara Puttaswany, Nguyen, Braverman, Hu and Liu, 2014; Vieira, Silva, Gigonzac, Ferreira, Gonalves and da Cruz, 2013). When the data set contains anomalous values, the adoption of the aforementioned methods will have two consequences: One is that correct result cannot be obtained through the multivariable estimation; and the second is that, the model residual error obtained according to the fitting is unable to detect all the anomalous values (Zhang, Li, Zhang, Wang, Zhao, Chen and Xu, 2012). However, the second consequence is rarely noticed by people. It is generally believed that, through the traditional fitting method such as OLS to carry out fitting on data, the anomalous values in the data can be detected in accordance with some classic statistics (Keimel, Habigt, Horch and Diepold, 2012). However, the diagnostic statistics estimated based on these traditional OLS estimates is primarily for a single abnormal point, if there are multiple abnormal points, the selection of these diagnostic methods may result in two consequences: One is the occurrence of the masking phenomenon, namely, the failure to identify some real abnormal points; The other is the swamping phenomenon, namely, the normal point is mistakenly determined as abnormal point, in addition, the influence of the masking phenomenon is particularly severe (Rocha, 2012; Cebi, 2013; Tingman, Jian and Zhang, 2013; Kim, Iwamoto, Kuffner, Ou and Pollard, 2013).
Robust statistical methods can not only produce estimation results that are less affected by anomalous values, but also have less bias in the fitting residual error, which can provide the information about anomalous values more prominently, and thus can better identify the anomalous values (Heradio, Fernández-Amorós, Cabrerizo and Herrera-Vizema, 2012; Han, Zhao, Xu, Peng, Chen and Li, 2013; Gao, Fu, Wei, Chen, Huang, Wang, Anderson and Wong, 2013). This paper has constructed the Combined Forecasting Model based on the minimum error absolute value sum. By the application of the model, GDP data is forecasted, and for the first time, this method is applied in the reliability evaluation of China's GDP data.

2. COMBINED FORECASTING MODEL AND DATA QUALITY EVALUATION METHOD

Different prediction methods may often provide different result based on the same information, and if simply discarding some methods that have relatively large error, it will result in the dropping of some useful information, which may make the accuracy of the model not high. The combination forecasting method is through the establishment of a combined forecasting model, to synthesize the predicted results obtained by a variety of prediction methods. As the combined model can make use of the variety of predictionsample information to a greater extent, it is more systematic and comprehensive in the consideration of problems than single forecasting model, and thus can effectively reduce the influence of random factors on the single forecasting model, so as to improve the accuracy and stability of the prediction.

2.1. Combined Forecasting Model

This paper tries to adopt the predictive value to replace the "True value" to perform data quality evaluation, to improve the prediction accuracy of the models as much as possible, the weight that can ensure the highest accuracy is therefore selected, so as to build the Combined Forecasting Model based on the minimum error absolute value sum.

Let \( \{y_i\} \) be a certain observation sequence, \( \{y_i(i)\} \) is the fitting sequence obtained by m models, \( w_i \) is the weight coefficient of the ith model in the Combined Forecasting Model, and \( w_i \) satisfies the constraint condition \( \sum_{i=1}^{m} w_i = 1, w_i \geq 0 \), then the prediction error \( e_t (t = 1,2,\ldots,n) \) of the Combined Forecasting Model is as the following: \( e_t = y_t - \sum_{i=1}^{m} w_i (y_i(i)) = \sum_{i=1}^{m} w_i \left( y_t - y_i(i) \right) \). Therefore, the absolute value of the fitting error of the Combined Forecasting Model can be expressed as the following: \( Q = \sum_{t=1}^{n} |e_t| \), and with the optimum of the minimum absolute value sum of the combined forecasting model as the principle, the mathematical model to determine the weight coefficient of m kinds of single forecasting methods is as the following:

\[
\min Q, Q = \sum_{t=1}^{n} |e_t| \\
\begin{align*}
\sum_{i=1}^{m} w_i &= 1 \\
w_i &\geq 0, i = 1,2,\ldots,m
\end{align*}
\]

(1)

The objective function in the aforementioned model contains absolute value, and it is not convenient to have a direct solution, but appropriate transformation can be made, so as to convert it into linear programming problem for the solution. Let \( u_t = \frac{|e_t| + e_t}{2}, v_t = \frac{|e_t| - e_t}{2} \), then obviously \( |e_t| = u_t + v_t, e_t = u_t - v_t, u_t \geq 0, v_t \geq 0 \), then the weight coefficient model is transformed into the linear programming problem as the following:

\[
\min Q = \min \left\{ \sum_{t=1}^{n} (u_t + v_t) \right\}
\]
The linear programming problem has \( m + 2 \) unknowns and \( n + 1 \) constraint, which can be solved by the adoption of the simple line method to get the optimal solution to the linear problem.

### 2.2. Data Quality Evaluation

After the specific evaluation model is estimated, for the choice of data quality evaluation method, this article from the perspective of the test on the anomalous values, analyze the prediction error, and finds out the anomalous values, by the application of Grubbs criteria and Dixon criteria to make inspection on the abnormal values of the relative error of the forecasting model, and to perform data quality evaluation. It is worth noting that the premise of the discussion on the anomalous values is that the main bodies of all the observation samples (the majority of the sample values except individual outliers) are from the same normal population or approximate normal population, therefore, it is necessary to perform normality test on the observation samples in advance.

1. **Grubbs Criteria**

   For \( n \) relative error data that is subject to normal distribution, after sorted from the smallest to the largest, calculate the test statistics \( G_n = \frac{P_{(n)}}{P_{(1)}}, G_i = \frac{\overline{P} - P_{(i)}}{S} \), in which \( P_{(n)} \) and \( P_{(1)} \) are the maximum and the minimum value respectively after sorting, and \( \overline{P} \) and \( S \) are the sample mean and sample standard deviation respectively. In the bilateral case, with the given significance level \( \alpha \), if \( G_n \geq G(\alpha, n) \), then \( P_{(n)} \) is considered as the abnormal value; If \( G_i \geq G(\alpha, n) \), then \( P_{(i)} \) is considered as the abnormal value.

   Where \( G(\alpha, n) \) is the critical value of the test statistics, it can be obtained through the table look-up.

2. **Dixon Criteria**

   For \( n \) relative error data \( P_1, P_2, P_3, \ldots, P_n \) that is subject to the normal distribution, after sorted from the smallest to the largest, the order statistics \( P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(n)} \) can be obtained. The test statistics is selected according to the sample size \( n \) as well as the location and number of the suspicious values, which details are given in Table 1:

<table>
<thead>
<tr>
<th>Sample Size n</th>
<th>High Value Anomaly Test Statistics D</th>
<th>Low Value Anomaly Test Statistics D</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-7</td>
<td>( D = \frac{P_{(n)} - P_{(n-1)}}{P_{(n)} - P_{(1)}} )</td>
<td>( D' = \frac{P_{(2)} - P_{(1)}}{P_{(n)} - P_{(1)}} )</td>
</tr>
<tr>
<td>8-10</td>
<td>( D = \frac{P_{(n)} - P_{(n-1)}}{P_{(n)} - P_{(2)}} )</td>
<td>( D' = \frac{P_{(2)} - P_{(1)}}{P_{(n-1)} - P_{(1)}} )</td>
</tr>
<tr>
<td>11-13</td>
<td>( D = \frac{P_{(n)} - P_{(n-2)}}{P_{(n)} - P_{(2)}} )</td>
<td>( D' = \frac{P_{(3)} - P_{(1)}}{P_{(n-1)} - P_{(1)}} )</td>
</tr>
<tr>
<td>14-30</td>
<td>( D = \frac{P_{(n)} - P_{(n-2)}}{P_{(n)} - P_{(1)}} )</td>
<td>( D' = \frac{P_{(3)} - P_{(1)}}{P_{(n-2)} - P_{(1)}} )</td>
</tr>
</tbody>
</table>

In the bilateral case, given the significance level \( \alpha \), check the critical value \( D(\alpha, n) \) corresponding to \( n \) in the Dixon critical value table. If \( D > D', D > D(\alpha, n) \), it can be determined that \( P_{(n)} \) is the anomalous value; if \( D' > D, D' > D(\alpha, n) \), it can be determined that \( P_{(1)} \) is the anomalous value; Otherwise, it can be determined...
that there is no anomalous value in the data. According to mathematical proof, if there is only one anomalous value in a set of data, Grubbs criteria is better than that the Dxiion criterion, while when there is more than one abnormal value, the Dxiion criterion is better than Grubbs criteria.

3. EMPIRICAL ANALYSIS ON THE STATISTICAL DATA VALIDATION OF THE COMBINED FORECASTING MODEL

3.1. Accuracy Verification of GDP Data of China

This paper applies the sample set as both the training set and the evaluation set, which requires to make the assumption that the data of the sample set is basically credible, therefore, the GDP data from 1990 to 2015 is selected as the sample to carry out analysis, in order to eliminate the influence of the price factor, unified conversion of the data into the GDP at constant price in 1990 is carried out.

3.1.1. Empirical Analysis on the Single Grey Forecasting Model

The GDP data sequence at constant price from 1990 to 2015 in China is \( X^{(0)}(k) \), the sequence generated by its accumulation is \( X^{(1)}(k) \), and the differential equation of the \( GM(1,1) \) model is:

\[
\frac{dX^{(1)}(t)}{dt} + aX^{(1)}(t) = \mu
\]

Let \( \hat{a} \) be the parameter vector to be estimated, \( \hat{a} = \left( \begin{array}{c} a \\ \mu \end{array} \right) \), make use of the least squares for the solution, and obtain \( \mu = 7739.052, a = -0.097 \). Solve the differential equations, and then obtain the forecasting model:

\[
X^{(1)}(k) = \left[ X^{(0)}(1) - \frac{\mu}{a} \right] e^{-\frac{\mu}{a}} + \frac{\mu}{a} , (k0,1,2,..,25)
\]

88698.71709779682680, (k0,1,2,..25)

\[
X^{(0)}(k)X^{(1)}(k)X^{(1)}(k)
\]

(3)

The established model cannot be directly applied in the prediction, which needs to perform the gray prediction inspection, with the inspection results shown in Table 2:

<table>
<thead>
<tr>
<th>Table 2.</th>
<th>Grey Prediction Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Residual Error Test</td>
</tr>
<tr>
<td>max</td>
<td>φ(i)</td>
</tr>
<tr>
<td>9.982</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: max|φ(i)| is the maximum absolute relative error; min|φ(i)| is the minimum absolute relative error; and aver|φ(i)| is the mean absolute relative error.

As can be known from the residual error test results, the model has passed the correlation test and the posterior variance test: In the residual error test, the mean absolute relative error is 4.011%, which is less than 5%, however, max|φ(i)| > 5%, hence it is necessary to further modify the grey forecasting model, make use of residual error \( GM(1,1) \) to make modification on the model, and use the model residual error to establish residual error \( GM(1,1) \), the modeling steps are the same as \( GM(1,1) \), it can be obtained that \( \mu = 3987.778, a = -0.075 \), and the residual error \( GM(1,1) \) model is as the following:

\[
\hat{e}^{(1)}(k + 1) = 52962.031exp(0.075k) - 52183.287
\]

\[
\hat{e}'(k + 1) = \left[ \hat{e}^{(1)}(k + 1) \right] = 3972.1523exp(0.075(k - 1))
\]

(4)

On the basis of \( \hat{X}^{(1)}(k + 1) \) plus the correction term \( \hat{e}'(k + 1) \), the error correction model can be obtained as the following:

\[
\hat{X}^{(1)}(k + 1) = 88698.71exp(0.097k) - 79682.680 + \delta(k - 1) \left( 3972.1523exp(0.075(k - 1)) \right)
\]

Wherein,

\[
\delta(k - 1) = \begin{cases} 1, k \geq 2, \\ 2, k < 2 \end{cases}
\]

and finally the original sequence forecasting model through residual error correction is given as the following:
\[ \hat{X}^{(i)}(k+1) = \hat{X}^{(i)}(k+1) - \hat{X}^{(i)}(k) (k = 1, 2, \ldots) \]  

(5)

3.1.2 Empirical Analysis on the Combined Forecasting Model

Make use of the model to make forecast on the GDP data from 1990 to 2015, and the GDP forecast sequence \( \{y_1(1)\}, \{y_1(2)\} \) and \( \{y_1(3)\} \) can be obtained. Run the lingo software for solution, and the optimal weight that makes the error absolute value sum of the combined forecasting model minimum can be obtained as the following:

\[ w_1 = 0.09, w_2 = 0.08, w_3 = 0.83 \]

Then the combined forecasting model is:

\[ y_t = w_1y_{1t} + w_2y_{2t} + w_3y_{3t} = 0.09y_{1t} + 0.08y_{2t} + 0.83y_{3t} \]

In order to test the effectiveness of the combined forecasting model, the accuracy of the established three single forecasting models and the combined forecasting model is compared, and the mean relative absolute error MAPE is selected for the evaluation indicator, with the calculation results shown in Table 3:

Table 3. Model Prediction Accuracy Evaluation Results

<table>
<thead>
<tr>
<th>Evaluation Indicator</th>
<th>GM (1,1) Model</th>
<th>Regression Combined Model</th>
<th>Double Exponential Smoothing</th>
<th>Combined Forecasting Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPE(%)</td>
<td>3.49</td>
<td>2.81</td>
<td>1.67</td>
<td>1.26</td>
</tr>
</tbody>
</table>

As can be known from Table 3, the MAPE of the combined forecasting model is the smallest, and the prediction accuracy is the highest, and the forecasting value obtained by the application of this model on the prediction of the GDP data from 1990 to 2015 in China is closer to the "True Value", and the reliability of the data quality evaluation is higher.

3.1.3 Data Quality Evaluation

According to the combined forecasting model to calculate the relative error of GDP data between 1990 and 2015 as the following:

\[ P_i = \frac{GDP_i - \hat{GDP}_i}{GDP_i} \times 100\% \]  

(6)

\[ \max \{P_i\} = 3.33\%, \min \{P_i\} = -5.52\%, \frac{1}{26} \sum_{i=1}^{26} |P_i| = 1.26\% \] is obtained, in which, except that the absolute value of the relative error of the GDP data in 1990 is greater than 5%, the absolute values of the relative error for the rest of years are all within 5%, it can be considered that the combined forecasting model is stable, which is a relatively real depiction of the GDP sequence, and can provide basic data for the evaluation of data quality.

Before testing the anomalous values of the relative error sequence \( \{P_i\} \), the normality test is carried out, and the K - S single sample distribution test is adopted. Run the SPSS software, and obtain the K - 1.216 S statistics as 1.216, the concomitant probability is 0.104, which is greater than 0.05, and the null hypothesis is accepted, namely, it can be considered that the sequence \( \{P_i\} \) is approximately subject to the normal distribution, which meets the applicable condition of the significance test of the abnormal values.

If there are anomalous values existed in the relative error sequence \( \{P_i\} \), they are most likely to occur at the ends of the sequence \( \{P_i\} \), and the outliers may be high end values, and may also be low end values, therefore, both the Grubbs inspection and Dixon inspection take the bilateral case into consideration. Under the significance level of 5%, when the Grubbs criteria is applied for judgment, \( G_n = 1.705 \), \( G_n = 2.968 \), check the Grubbs inspection critical value table, it can be known that \( G(a, n) = 2.841, G_n > G_n \) and \( G_n > G(a, n) \), then \( P_{(i)} \) is the outlier ( \( P_{(i)} \) is corresponding to the GDP error percentage in 1990). Continue the test after excluding the abnormal value, and no other abnormal values can be found, which shows that by the application of the Grubbs criteria, only one significance anomalous value \( P_{(i)} \) is found.

Next make use of the Dixon criteria for the judgment, and it is still under the significance level of 5%, when the sample size is 26, \( D_n = \frac{P_{(2)} - P_{(2-2)}}{P_{(2)} - P_{(2-2)}} = 0.043 \), \( D_n = \frac{P_{(2)} - P_{(2-2)}}{P_{(2)} - P_{(2-2)}} = 0.352 \), look up in the table to obtain
the critical value $D(a,n)=0.436$, $D(n)>D_n$ and $D(n)>D(a,n)$, and no anomalous values are found. Therefore, only the GDP data in 1990 is outlier, which has the possibility of distortion, while the GDP data in the rest of the years is accurate.

3.2 Urban Water Consumption Statistical Data Validation

Select the amount of water consumption of a certain city in 2000 ~ 2006 as the data sample, and verify the application of the gray oscillation sequence combined forecasting model in the forecast of urban water consumption. For the specific data please refer to Table 4.

Table 4. Raw Data of Urban Water Consumption of a Certain City from 2000 to 2006

<table>
<thead>
<tr>
<th>Year</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban Water Consumption (billion cubic meters)</td>
<td>10.35</td>
<td>10.89</td>
<td>9.52</td>
<td>9.04</td>
<td>9.45</td>
<td>10.28</td>
<td>11.97</td>
</tr>
</tbody>
</table>

Take the aforementioned annual water consumption data of a certain city from 2000 to 2006 as the original data sequence, due to the effect of various environmental factors, the urban water consumption statistics is a random oscillation sequence. Take the annual water consumption data from 2000 to 2004 as the original data sequence, and apply the MATLAB to prepare the computation program respectively and complete the GM (1, 1) model, Verhust model and combined forecasting model, complete the model construction, predict the urban water consumption statistical data in 2005 and 2006, and make evaluation on the prediction accuracy of the models.

Adopt the annual water consumption data from 2000 to 2004 as the original data sequence, and construct the combined forecasting model. Generate an accumulative sequence and mean generation sequence data with consecutive neighbors, and calculate power exponent $r$ of the model.

$$r = \frac{1}{n-2} \sum_{i=2}^{n} r_i = \frac{1}{3}(0.5985 - 0.8262 - 2.1930) = -0.8069$$

Apply the least square estimation and calculate:

$$\beta = (a, b)^T = (-0.1253, 83.508)^T$$

The time response sequence of the combination forecasting model is as the following

$$x_{k+1}^{(1)} = \left\{ \frac{b}{a} + \left[ \left( x_1^{(0)} - \frac{b}{a} \right) e^{(1-r)a} \right] \right\}^{\frac{1}{r}} \tag{7}$$

Table 5. Comparison of the Prediction Accuracy of the Models

<table>
<thead>
<tr>
<th>Year</th>
<th>Statistical Value</th>
<th>GM(1,1) Model</th>
<th>Verhust Model</th>
<th>GM(1,1) Power Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predictive Value</td>
<td>Relative Error</td>
<td>Predictive Value</td>
<td>Relative Error</td>
</tr>
<tr>
<td>2005</td>
<td>10.28</td>
<td>8.54</td>
<td>16.9%</td>
<td>9.00</td>
</tr>
<tr>
<td>2006</td>
<td>11.97</td>
<td>8.11</td>
<td>32.2%</td>
<td>8.78</td>
</tr>
</tbody>
</table>

Table 6. Comparison of the Prediction Accuracy of the Change of the Power Exponent $r$ of the Combined Forecasting Model

<table>
<thead>
<tr>
<th>Year</th>
<th>Statistical Value</th>
<th>$r=r$</th>
<th>$r=r-0.002$</th>
<th>$r=r-0.002$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predictive Value</td>
<td>Relative Error</td>
<td>Predictive Value</td>
<td>Relative Error</td>
</tr>
<tr>
<td>2005</td>
<td>10.28</td>
<td>11.03</td>
<td>7.2%</td>
<td>10.98</td>
</tr>
<tr>
<td>2006</td>
<td>11.97</td>
<td>12.12</td>
<td>1.2%</td>
<td>12.07</td>
</tr>
</tbody>
</table>

As can be seen from the data comparison in Table 5, the accuracy of the combined forecasting model is significantly higher than that of the GM (1, 1) model and Verhust model; And it can be seen from the comparison data of the forecasting accuracy of the changes in the power exponent $r$ of the combined forecasting model that, the changes of the power exponent can improve the prediction accuracy, but at this point, attention shall be paid to the change of the fitting precision of the model.

4. CONCLUSION

Make use of the statistical model to evaluate the data quality, the basic idea is to apply the predicted values of the model as the “True Value”, and then check whether the difference of the predicted values and the
evaluation data is significant, if the difference is significant, abnormal values exist in the data, and if the abnormal value is not objective reason (such as external impact, institutional reform, etc.), it can be regarded as the accuracy problem of the data itself. However, there is certain difference between the predicted values obtained in accordance with the historical data and the real values, in order to get a more accurate conclusion, it is necessary to adopt a model with better forecasting effect to carry out prediction, so as to narrow the differences as much as possible. The combined forecasting model can incorporate the single forecasting models of diverse nature, and concentrate more economic information and prediction technology, reduce the systematic error of the prediction system, so as to significantly improve the prediction effect. It can be seen that, compared with the single forecasting model, the application of the combined forecasting model to test the accuracy of the data will have better effect of the inspection. This paper is based on the accuracy of the Chinese GDP data from 1990 to 2015 as an example for the empirical analysis, in accordance with the characteristics of the examining indicators to establish three single forecasting models including the grey forecasting model, regression combined model and double exponential smoothing model, and then according to the basic idea of the combined forecasting model, carry out weighted combination on predictive values of the three single forecasting models to obtain the combined forecasting model. Through the comparison of the prediction precision of each single forecasting model and the combined forecasting model, it can be found that the combined forecasting model is better than the single forecasting model, and the predicted value obtained is closer to the "True Value", which is thus more suitable for the accuracy test of the data.

REFERENCES


