Application of grey relational analysis for multivariate time series

Roselina Sallehuddin, Siti Mariyam Hj. Shamsuddin, Siti Zaiton Mohd Hashim,
Fakulti Sains Komputer dan Sistem Maklumat,
Universiti Teknologi Malaysia, 81300, Skudai, Johor
roselina, marilyam, sitizaiton @utm.my

Abstract

Grey Relational Analysis (GRA) has been widely applied in analysing Multivariate Time Series data (MTS). It is an alternate solution to the traditional statistical limitations. GRA is employed to search for Grey Relational Grade (GRG) which can be used to describe the relationships between the data attributes and to determine the important factors that significantly influence some defined objectives. This paper demonstrates how GRA has been successfully used in identifying the significant factors that affect the grain crop yield in China from 1990 to 2003. The results from analysing the sample data revealed that the main factor that affects the trend of crop yield is the consumption of pesticide and chemical fertilizer and the least important factor to be considered is the agricultural labour. Thus, by properly adjusting the significant affecting factors, the China’s crop yield performance can be further improved. Furthermore, GRA can provide a ranking scheme that gives the order of the grey relationship among the dependent and independent factors which leads to essential information such as which input factor need to be considered to forecast grain crop yield more precisely when using Artificial Neural Network (ANN). In order to evaluate the performance of GRA in ANN model, a comparison is made using Multiple Linear Regression (MR) statistical method. The results from the experiment show that ANN using GRA has outperformed the MR model with 99.0% in forecasting accuracy.

Keywords: Grey Relational Analysis, Grey Relational Grade, Multiple Linear Regression, Multivariate, Artificial Neural Network

1. Introduction

In a complex and multivariate time series system, such as social and economic systems, there are many factors that can influence the state of the system. Getting information such as the most and the least significant factors that influence the system’s performance is very important. However, identifying the relationship among the contributing factors is always grey particularly when the information is not clear, incomplete and uncertain. It is also difficult to get the practical and experimental data. Factor analysis and regression analysis methods are two conventional multivariate statistical methods that frequently used to represent the relationship between independent and dependent factors. However, this analysis prescribes that there must be relationship of mutual influence between variables. The function relationship will only work if there is a large quantity of data which should conform to the typical distribution, like the normal distribution. However, in practice, some difficulties do exist. For instance, it is difficult to extract the significant financial ratio variables and financial indicators which affect the financial performance of venture capital in Taiwan due to lack of sample data [1]. As a result, a new method of analysis is needed.

To overcome the restrictions of regression analysis and factor analysis, a multi-attribute method, Grey Relational Analysis (GRA) has been proposed to solve the problem [1-3]. The GRA is an effective tool to make system analysis, and also lays groundwork for modeling, forecasting, and clustering of grey systems. Compared to regression analysis and factor analysis in mathematical statistics, GRA has some advantages such as requiring only small sample size, not using typical distribution, and not requiring independent factor and using simple calculation. The GRA analysis has been proven to be simple and accurate for selecting significant factors especially for problems with unique characteristic [4]. In this study, the GRA is utilized to establish a ranking scheme that ranks the order of the grey relationship among dependent and independent factors based on Grey Relational Grade (GRG) values. This scheme is known as an evaluation model for selecting significant factors in multivariate time series where
GRG is re-arranged according to the order of their magnitude.

The effectiveness of these selected factors can be evaluated using forecasting model such as a grey prediction model (GP) and Artificial Neural Networks model (ANNs). In this paper, the significant factors will be used as input node in ANNs forecasting model because it outperformed GP model [3]. For the case study, the grain crop yield of China and its affecting factors will be employed. GRA will find out the relationships among the affecting factors and then determine the most and the least important factors that significantly influence the trend of grain crop yield. Subsequently, GRA will recommend the number of input node for ANN forecasting model.

The rest of this paper is organized as the following. Section 2 describes the grey theory and grey relational analysis. Section 3 presents the GRA method. In Section 4, experimental data used to evaluate the GRA model is described. In section 5, the experimental results obtained from GRA are discussed, and finally section 6 provides the concluding remarks.

2. Grey Theory and Grey Relational Analysis

Grey System (GS) is the system of which part of the information is known and the other part of information is unknown. Up to now, GS theory has developed a set of theories and techniques including grey mathematics, grey relational analysis, grey modelling, grey clustering, grey forecasting, grey decision making, grey programming and grey control which has been applied successfully in many engineering and managerial fields such as industry, ecology, meteorology, geography, earthquake, hydrology, medicine and military [5-7]. The major advantage of grey theory is that it can handle both incomplete information and unclear problems very precisely. It serves as an analyzing tool especially in cases where there is insufficient data [3].

GRA is a new method of analysis which has been proposed in the Grey system theory and it is founded by Professor Deng [4]. GRA is based on geometrical mathematics which complies with the principles of normality, symmetry, entirety, and proximity. GRA is suitable for solving complicated inter-relationships between multiple factors and variables and has been successfully applied on cluster analysis, robot path planning, project selection, prediction analysis, performance evaluation, and factor-effect evaluation and multiple criteria decision [7-10]. Detailed explanation about GRA method is presented in the following section.

3. Grey Relational Analysis Method

There are three (3) main steps in GRA. The first step is data pre-processing. Data pre-processing is usually required when the range or unit in one data sequence is different from others or the sequence scatter range is too large. Here, the data must be normalized, scaled and polarized into a comparable sequence before proceeding with the other steps. The process is called the generation of grey relation or standard processing.

There are two processes involved in pre-processing stage; data representative and data normalization. Initially the original data series \( X \) is represented as reference \( x_0 \) and comparative series \( x_i \). In crop dataset, \( x_0 \) represents the China’s crop yield and \( x_i \) represents the ten affecting factors that influence the production of China’s crop. Next, the data is normalized. There are several ways of doing data pre-processing for the GRA such as (1) and (2), [1, 13]. To determine which one to be used is based on the characteristics of data sequence, for example:

If the expectancy is the higher-the-better, it can be expressed as

\[
x^* (k) = \frac{x_0^0 (k) - \min x_i^0 (k)}{\max x_i^0 (k) - \min x_i^0 (k)}
\]  

(1)

If the expectancy is the lower-the-better, it can be expressed as

\[
x^* (k) = \frac{\max x_i^0 (k) - x_i^0 (k)}{\max x_i^0 (k) - \min x_i^0 (k)}
\]

(2)

where,

\( i = 1, ..., m; \quad k = 1, ..., n \).

\( m \) is number of experimental data items,

\( n \) is the number of parameters;

\( x_i^0 (k) \) is the original sequence,

\( x_i^* (k) \) is the sequences after data pre processing,

\( \min x_i^0 (k) \) and \( \max x_i^0 (k) \) are the smallest and the largest value of \( x_i^0 (k) \).

In this study, (1) is employed, since the output of this study has the characteristic of “higher is better”, meaning if the value of GRG is higher, than there is a strong relationship between comparative and
reference series. The range of data is adjusted so as to fall within [0,1] range.

The second step is to locate the grey relational co-efficient by using (3), [8]:

\[
\xi_i(k) = \frac{\Delta \min + \zeta \Delta \max}{\Delta_{0,j}(k) + \zeta \Delta \max}
\]

(3)

where,

\[
\Delta_{0,j} = \|x_0(k) - x_j^*(k)\|
\]

\[
\Delta \min = \min_{\forall j \in V_k} \|x_0(k) - x_j^*(k)\|
\]

\[
\Delta \max = \max_{\forall j \in V_k} \|x_0(k) - x_j^*(k)\|
\]

\[
x_0^*(k) = \text{the reference sequence, and} \]

\[
x_j^*(k) = \text{the comparative sequence.}
\]

\(\zeta\) is known as an identification co-efficient with \(\zeta \in [0,1]\) which can be adjusted to have better distinction between normalized reference series and normalized comparative series. Normally \(\zeta =0.5\) is used because it offers moderate distinguishing effect and stability [7]. Furthermore, based on mathematical proof, the value change of \(\zeta\) will only change the magnitude of the relational co-efficient but it would not change the rank of the grey relational grade [11].

From the calculation, the values of \(\Delta \min\) and \(\Delta \max\) are respectively 0 and 1. By replacing these values in (3), we obtained:

\[
\xi_i(k) = \frac{0 + \zeta(l)}{\Delta_{0,j}(k) + \zeta(l)} = \frac{0.5}{\Delta_{0,j}(k) + 0.5}
\]

(4)

After the grey relational co-efficient is derived, grey relational grade (GRG) is calculated by averaging the value of the grey relational co-efficient [8-13]. GRG is defined as the numerical measure of the relevance between two systems or two sequences such as the reference sequence and the comparative sequence. The existing GRG between two series is always distributed between 0 and 1. GRG can be calculated using (5) [8-13]:

\[
\gamma_i = \frac{1}{n} \sum_{k=1}^{n} \xi_i(k)
\]

(5)

where \(\gamma_i\) represents GRG; the level of correlation between the reference sequence and the comparative sequence. In this study, GRG is used to indicate the degree of influence that the comparative sequence (10 affecting factors) could exert over the reference sequence (China’s crop yield). Therefore, if a particular comparative sequence is more important than the reference sequence, then the GRG for that comparative sequences and reference sequence will be higher than other GRG [11]. For example, if \(\gamma(x_0, x_i) > \gamma(x_0, x_j)\), then the element \(x_i\) is closer to the reference element \(x_0\) than the element \(x_j\). Generally, \(\gamma_i > 0.9\) indicates a marked influence, \(\gamma_i > 0.8\) a relatively marked influence, \(\gamma_i > 0.7\) a noticeable influence and \(\gamma_i < 0.6\) a negligible influence [12].

4. Experimental Data Setup

To facilitate this study, a dataset obtained from [4] is used as experimental data. This sample contains 13 observations specifying annual China gross grain crop yields and their affecting factors. Due limited pages, the original data is not listed here, and the detailed can be obtained from [4] or China Statistical Yearbook (1991-2003) or China Agriculture Statistical Yearbook (1991-2003).

Ten factors that affect the production of gross grain crop in China include \((a)\) total power of agricultural, \((b)\) electricity consumed in rural areas, \((c)\) irrigation area, \((e)\) consumption of chemical fertilizer, \((f)\) areas affected by natural disaster, \((g)\) budgetary expenditure for agriculture, \((h)\) sown area of grain crops, \((l)\) consumption of pesticide, \((m)\) consumption of agricultural film, and \((n)\) agriculture labourers. The total production of grain crop yield is denoted by \((d)\).

In this study, the data is divided into two parts. First part (1991-2001) is used to construct the evaluation model and to train the ANN. Data from 2002 to 2003 is used to validate the performance of ANN forecasting model where forecasting one step ahead is implemented.
5. Results and discussion

In GRA evaluation model, experimental data are first normalized in the range between zero and one, which is called grey relational generation. Subsequently, the grey relational co-efficient is calculated from the normalized experimental data to express the relationship between the desired and actual experimental data. The GRG is then computed by averaging the grey relational co-efficient corresponds to each response. The overall evaluation of the multiple process response is based on the GRG.

This section will discuss on the final result given by GRA analysis. Table 1 demonstrates the calculated value of GRG for each affecting factors \( x_i \) used in this study to China’s crop yield \( x_0 \). Table 1 shows that GRG values in the range of [0, 1] and each affecting factor and crop yield with positive correlation. These results indicate that each affecting factors has influenced in the production of China grain crop.

| Table 1: Grey Relational Grade for each affecting factor (Non-ordered) |
|--------------------------|----------|
| \( x_i \) | GRG orders | \( x_0 \) | GRG orders |
| m | 0.788 | 7 | g | 0.741 | 8 |
| b | 0.801 | 5 | h | 0.809 | 3 |
| c | 0.803 | 4 | l | 0.830 | 1 |
| e | 0.822 | 2 | a | 0.800 | 6 |
| f | 0.737 | 9 | n | 0.717 | 10 |

Based on the calculated value of GRG, the grey relational order of each affecting factors can be determined. Thus, a ranking scheme based on the size of \( \gamma_i \) is constructed as shown in Table 2 and Figure 1. Each \( \gamma_i \) is ordered to the increasing grey relational co-efficient. This derived order gives the priority list in choosing the affecting factors that are closely related to the crop yield, \( x_o \). Since the GRG represent the level of correlation between the crop yield sequence and affecting factors sequence, the greater value of GRG means that the affecting factor sequence has a stronger correlation to the crop yield sequence [12]. Therefore, from Table 1, we can rank the values of \( \gamma(x_0,x_i) \) as in Table 2 or as ranking scheme in Figure 1.

Thus, based on the GRG values that present the relevance of each affecting factors and crop yield, we can say that the consumption of pesticide \( l \) affects the crop production the most, followed by the consumption of chemical fertilizer \( e \) and sown area of grain crops \( h \).

| Table 2: Grey relational grade for each affecting factors (Ordered sequence) |
|--------------------------|----------|
| \( x_i \) | GRG orders | \( x_0 \) | GRG orders |
| l | 0.830 | 1 | a | 0.800 | 6 |
| c | 0.822 | 2 | m | 0.788 | 7 |
| h | 0.809 | 3 | g | 0.741 | 8 |
| e | 0.803 | 4 | f | 0.737 | 9 |
| b | 0.801 | 5 | n | 0.717 | 10 |

Subsequently, the irrigation area \( c \), electricity consumed in rural areas \( h \), consumption of agricultural film \( m \), total power of agricultural \( a \), and budgetary expenditure for agriculture \( g \) gave the moderate influence to the crop yield. While, the areas affected by natural disaster \( f \), and agriculture labourers \( n \) gave the least effect to the pattern of crop yield.

Figure 1. A ranking scheme for China Crop Yield

Hence the China agriculture department should focus their attention more on these influential factors in order to monitor the performance of China grain crop production more effectively. These influential factors also can be employed in forecasting model for more accurate result. On the contrary, the forecaster can pay less attention to the less influential factors as these elements contribute relatively little to the variation of grain crop yield. From the experiment, it can be concluded that all the input with GRG less than 0.8 can be omitted and labelled as least important. As a result, the application of GRA has reduced the number of affecting factors that affect the production of China grain crop from 10 to 6.

Based on the calculated value of GRG, only six factors: \( a, b, c, e, h \) and \( l \) are selected as the inputs to ANN to predict the grain crop yield for year 2002 and 2003. By reducing the input number, GRA also help to simplify the network structure and consequently speed up the network learning time. The network structure used to train ANN forecasting model is 6:12:1, where 6 represent six input nodes (six affecting factors: \( a, b, c, e, h \) and \( l \)), 12 represents twelve hidden nodes and 1 represent one output node; in this case, grain crop yield. The learning parameters used in this study are 0.5 and 0.9, which represent the learning rate and the momentum value, respectively.
In order to evaluate the performance of GRA in selecting significant input factors for ANN model, a comparison with traditional statistical model, multiple linear regressions (MR), is carried out. In MR analysis, goodness fit test is used to recognize the proper inputs. The significant inputs are identified based on t-values and p-values. If t-values are less than 1 and p-values are above some accepted level, such as 0.05, then this variable is excluded from the list. Initially, the ten independent variables are used to build MR model for grain crop yield. Several models are built and evaluated based on statistical goodness fit. However, the final model only uses 4 independent variables; $a, c, h$ and $n$.

The equation for grain crop yield is given as:

$$\text{Crop yield}_t = f(a_t, c_t, h_t, n_t)$$

$$= -2.45a_t + 2.83c_t + 0.44h_t - 0.61n_t - 92298$$

(6)

where,

- $a_t$ is a total power of agricultural at period $t$,
- $c_t$ is an irrigation area at period $t$,
- $h_t$ is a sown area of grain crops at period $t$, and
- $n_t$ is an agriculture labours at period $t$.

Equation 6 shows that the most influencing factors for China’s crop yield is sown area of grain crops ($h$), followed by total power of agricultural ($a$), then agriculture labours ($n$) and an irrigation area ($c$). An irrigation area and sown area have positive correlation with crop yield. These indicate that both affecting factors have influenced on the production of China grain crop. Meanwhile, the total power of agricultural and agriculture labours have negative correlation and these indicate that both factors influence crop yield in a negative direction.

Table 3: Forecasting performance

<table>
<thead>
<tr>
<th>Predicted Year</th>
<th>Actual data</th>
<th>ANN</th>
<th>MR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>45263.70</td>
<td>45088.35</td>
<td>45376.4</td>
</tr>
<tr>
<td>2003</td>
<td>45705.80</td>
<td>45142.26</td>
<td>43175.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RMSE</th>
<th>%accuracy</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>417.33</td>
<td>99%</td>
<td>1790.96</td>
<td>97%</td>
</tr>
</tbody>
</table>

Table 3 and Figure 2 illustrate the forecasting values produced by ANN model and MR model. Root mean square error (RMSE) is used to evaluate the forecasting performance for ANN and MR. In this experiment, RMSE for ANN and MR models are 417.33 and 1790.96 with 99% and 97% accuracy performance respectively.

From Table 3 and Figure 2, the production of China crop yield has increased about 0.90% from year 2002 to year 2003. Meanwhile, ANN forecasting model prediction shows there is a slight increment of 0.20% from 2002 to 2003. However, contradicting result is obtained from MR model that shows the production of grain crop is decreased about 4.80% in 2003. Experimental results clearly show that the prediction given by ANN is more significant and reliable because it manages to produce smaller forecasting errors and can correctly predict the movement of grain crop yield.

The result indicates that compared to goodness fit in MR, GRA analysis in ANN is more accurate in finding the significant affecting factors. Therefore, GRA is recommended for analyzing multivariate time series data with small sample size and incomplete information.

6. Conclusion

GRA is part of grey system theory which is suitable for solving complicated inter-relationships between multiple factors and variables in MTS. In this case study, the affecting factors such as the consumption of pesticide, the consumption of chemical fertilizer, sown area of grain crops, the irrigation area, electricity consumed in rural areas, consumption of agricultural film, total power of agricultural, budgetary expenditure for agriculture, the areas affected by natural disaster, and agriculture laborers have an effect to the control variables; China’s crop yield. This approach is based on the level of similarity and variability among all factors to establish their relationship.

The outcome from the GRA as per Table 2 shows the order of importance of the affecting factors to the China’s grain crop yield is the consumption of pesticide ($l$), the consumption of chemical fertilizer ($e$), sown area of grain crops ($h$), the irrigation area ($c$), electricity consumed in rural areas ($b$),...
consumption of agricultural film \((m)\), total power of agricultural \((a)\), budgetary expenditure for agriculture \((g)\), the areas affected by natural disaster \((f)\), and agriculture labourers \((n)\). Experimental results have shown clearly that the consumption of pesticide \((l)\) is the most influential factor to China’s grain crop and the labour is the least influential factor that will affect the China’s crop production. The GRA of GRG also provides us with an alternative to decide which input factors that show the crucial effect to the crop productions. In this paper, only six input factors; the consumption of pesticide \((l)\), the consumption of chemical fertilizer \((e)\), sown area of grain crops \((h)\), the irrigation area \((c)\), electricity consumed in rural areas \((b)\), and consumption of agricultural film, are considered as the most influential factors. Then, they are chosen as input node to ANN forecasting value.

The results from this study also illustrates that compared to traditional statistical method, MR, GRA is better in analyzing relationship among affecting factors and crop yield. Moreover, ANN model using GRA is also able to give more accurate and reliable forecasting values. Therefore, it can be concluded that GRA gives a favorable tool as analyzing that compared to small and incomplete multivariate time series data. For future study, we attempt to extend the application of GRA in analyzing and predicting real time series problem such as predicting stocks closing price at Kuala Lumpur Stock Exchange.

Acknowledgement

Authors would like to thank Ministry of Higher Education for sponsoring this research under Fundamental Research Grant Scheme (FRGS - 78243), and Research Management Center for the support and research activities.

References


