Highlights

- A novel discrete grey polynomial model is proposed.
- The proposed model unifies the univariate discrete grey models.
- An algorithm is presented to select the optimal model structure adaptively.
- Matrix decomposition is adopted to provide a simpler paradigm for property analysis.

Data-based structure selection for unified discrete grey prediction model

Bao-lei Wei[a,*], Nai-ming Xie[a], Ying-jie Yang[b]

[a] *College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, China*
[b] *Institute of Artificial Intelligence, De Montfort University, Leicester, United Kingdom*

Abstract

Grey models have been reported to be promising for time series prediction with small samples, but the diversity kinds of model structures and modelling assumptions restrains their further applications and developments. In this paper, a novel grey prediction model, named discrete grey polynomial model, is proposed to unify a family of univariate discrete grey models. The proposed model has the capacity to represent most popular homogeneous and non-homogeneous discrete grey models and furthermore, it can induce some other novel models, thereby highlighting the relationship between the models and their structures and assumptions. Based on the proposed model, a data-based algorithm is put forward to select the model structure adaptively. It reduces the requirement for modeler's knowledge from an expert system perspective. Two numerical experiments with large-scale simulations are conducted and the results show its effectiveness. In the end, two real case tests show that the proposed model benefits from its adaptive structure and produces reliable multi-step ahead predictions.

Keywords: grey system theory, discrete grey model, structure selection, matrix decomposition

1. Introduction

Time series prediction has been an important topic in various fields ranging from natural science to social science. In the past decades, researchers have proposed many kinds of statistical and machine learning methods for large-sample time series prediction, among which the most representative techniques include the exponential smoothing, autoregressive integrated moving average, support vector regression, and neural network, etc (Box et al., 1994; Hastie et al., 2013; Tim et al., 1996). These methods work well when large data collection is available. However, sometimes it is difficult or even impossible to collect sufficient data, such as the cold-start problems in recommendation systems (L et al., 2012) and the incomplete information in disaster emergency managements (Altay & Iii, 2007). From an expert system perspective, the modelers are required to have a higher level of skill and knowledge to infer the likely consequences if only limited data are available, restraining the application of expert system in practice. Therefore, researchers have been paying attention to the identification of effective predicting models that only require small samples (Liu et al., 2017).

Grey system theory is one of the most popular techniques for small-sample time series analysis and prediction. Grey models use the accumulating generation sequence operator to mine the pattern hiding in the small-sample time series, and then quantify the identified pattern using dynamic equations instead of static formulas. In this way, the grey-based predictors can achieve high accuracy with smaller number of samples (Xie & Wang, 2017).

There are many different available grey models. Grey model GM(1,1), employing the first-order (the first '1') single-variable (the second '1') differential equation, is the first and also the most popular one.

*Corresponding author
Email addresses: weibaolei_2014@163.com (Bao-lei Wei), xienaiming@nuaa.edu.cn (Nai-ming Xie), yyang@dmu.ac.uk (Ying-jie Yang)

Since grey models assume that the original time series is quasi-exponential, and this is in agreement with the physics that the accumulation and release in generalized energy system including economy, society and ecosystem usually conform to exponential law (Liu et al., 2017). It has been extensively researched from the methodological and industrial viewpoints, such as the necessary and sufficient condition for modeling (Chen & Huang, 2013), relative error bound estimation (Liu et al., 2014), and influence of sample size on modeling performance (Yao et al., 2009; Wu et al., 2013a; Tabaszewski & Cempel, 2015). Meanwhile, some researchers proposed novel approaches to improve the modeling accuracy by generalizing the accumulating generation form integer-order to fraction-order (Wu et al., 2013b), optimizing the background coefficient optimization (Zhao et al., 2012) and initial value (Xu et al., 2011), and hybrid approaches optimizing them simultaneously (Wang et al., 2014; Xiao et al., 2014). However, in spite of their significant improvements on modeling accuracy, GM(1,1) is still limited to the quasi-exponential time series. Therefore, researchers inspired by the modeling ideas of GM(1,1), proposed a series of novel models to obtain wider application scopes. For example, the Bernoulli differential equation, NBGM(1,1) was firstly proposed for the time series with an inverted U-shaped trajectory (Chen et al., 2008; Wang et al., 2011; Evans, 2014; Shaikh et al., 2017). Subsequently, the non-homogeneous grey model, NGM(1,1), was developed for quasi non-homogeneous exponential series by replacing the constant term in the differential equation of GM(1,1) with a linear function of time; a generalization form of the non-homogeneous grey model, GM(1,1,t^α), was proposed by replacing the linear function with the α-order power function. Until recently, Luo & Wei (2017) and Wei et al. (2018) developed a grey polynomial model, GPM(1,1,N), by taking the N-order polynomial function as the forcing term in the differential equation. It has been proved that the grey polynomial model unifies the above two models and is applicable to the complex nonlinear time series with quasi-homogeneous, non-homogeneous or even fluctuating characteristics.

Although the aforementioned optimization and development of grey predicting models play an important role in improving modeling accuracy and expanding application scopes, they still have some weaknesses. As pointed out by Xie & Liu (2009), the numerical integration (known as background value) or numerical derivative (known as grey derivative) will introduce discretization errors, resulting in the gap between model values and actual values even for the pure exponential series without noise. In order to avoid discretization errors, Xie and Liu firstly constructed the discrete grey model DGM(1,1) (Xie & Liu, 2009), the discrete form of GM(1,1), and then the non-homogeneous discrete grey model NDGM(1,1) (Xie et al., 2013a), the discrete form of NGM(1,1), by using the individual difference equations directly. Based on this argument, researchers also proposed some other discrete models, such as the nonlinear ones (Long et al., 2014; Zhang et al., 2015), the fractional order accumulation-based ones (Wu et al., 2014), the multivariate ones (Ma & Liu, 2016; Wu et al., 2016) and the interval-based ones (Liu & Shyr, 2005; Zeng et al., 2010; Xie & Liu, 2015). By analyzing the modeling procedure, the existing discrete grey models have the following shortcomings. First, there are a variety of discrete grey models with different structures, which makes it difficult for researchers to provide a unified paradigm for property analysis. Next, even though each discrete grey model has explicit assumption, it is difficult to determine whether the data meets the modeling condition in practice exactly, leading to challenges for users to select an appropriate model. Last, the grey polynomial model were proved to have wider application scopes, but its discrete form has not yet been investigated systematically.

To overcome these problems, we proposed a novel discrete grey polynomial model and the main contributions can be summarized as follows:

(1) The proposed model unifies a family of univariate discrete grey models including the homogeneous and non-homogeneous ones, and can induce some other novel models. It fits the original time series directly rather than the accumulating generation series, and avoids the two-step parameter estimation and the unnecessary complexity introduced by the inverse accumulating generation. In particular, the unified representation makes it simple to provide a paradigm for property analysis by using the matrix decomposition.

(2) Different from the classical grey predicting models where the expressions are always preset, a data-based selection algorithm searches the optimal model structure adaptively. Large-scale simulations are conducted to statistically test the effectiveness and also the robustness, especially under the noisy environment. Two real data sets are used as benchmark to compare the proposed method with

existing methods in terms of their predicting accuracy, and the results show that the proposed method outperforms the alternatives.

The remainder of this paper has following structure: the proposed model and a data-based structure selection algorithm are presented in Sections 2 and 3; the theoretical property are analyzed in Section 4; large-scale simulations and two real case tests are provided in Sections 5 and 6; conclusions and future work are discussed in Section 7.

2. Discrete grey polynomial model

2.1. The definition of grey polynomial model

Supposing that the original time series is $X^{(0)} = \left\{x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)\right\}$, the accumulating generation series is defined as

$$X^{(1)} = \left\{x^{(1)}(1), x^{(1)}(2), \cdots, x^{(1)}(n)\right\},$$

where

$$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i),\ k = 1, 2, \cdots, n. \tag{1}$$

The grey polynomial prediction model (Luo & Wei, 2017) for the accumulating generation series $X^{(1)}$ is suggested to be the coupled equations composed of the differential equation

$$\frac{d}{dt}x^{(1)}(t) = ax^{(1)}(t) + b_0 + b_1 t + \cdots + b_N t^N, \tag{2}$$

and the difference equation

$$x^{(0)}(k) = a\left[(1-\lambda)x^{(1)}(k-1) + \lambda x^{(1)}(k)\right] + \sum_{j=0}^{N} b_j \frac{k^{j+1} - (k-1)^{j+1}}{j+1} + \varepsilon_k, \tag{3}$$

where $N < n-3$ is the polynomial order; $\lambda \in [0, 1]$ is the tuned background coefficient; ε_k is the discretization error. If the polynomial order N is equal to 0, the sample size n must be equal or greater than 4, which has always been the basic requirement of GM(1,1) model (Xie & Wang, 2017).

Based on the difference equation, the least square estimates of the model parameters $b_0, b_1, \cdots, b_N$ and a are calculated as

$$\begin{bmatrix}\hat{a} & \hat{b}_0 & \hat{b}_1 & \cdots & \hat{b}_N\end{bmatrix}^{\mathrm{T}} = \left(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{y}, \tag{4}$$

where

$$\boldsymbol{y} = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix},\ \boldsymbol{B} = \begin{bmatrix} (1-\lambda)x^{(1)}(1) + \lambda x^{(1)}(2) & 1 & \frac{3}{2} & \cdots & \frac{2^{N+1}-1^{N+1}}{N+1} \\ (1-\lambda)x^{(2)}(1) + \lambda x^{(1)}(3) & 1 & \frac{5}{2} & \cdots & \frac{3^{N+1}-2^{N+1}}{N+1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ (1-\lambda)x^{(1)}(n-1) + \lambda x^{(1)}(n) & 1 & \frac{2n-1}{2} & \cdots & \frac{n^{N+1}-(n-1)^{N+1}}{N+1} \end{bmatrix}.$$

Substituting the estimates into the analytic solution to the differential equation (2) (see the details in Luo & Wei (2017)) gives the time response function

$$\hat{x}^{(1)}(t) = ce^{\hat{a}t} + d_0 + d_1 t + \cdots + d_N t^N \tag{5}$$

where the polynomial coefficients $d_0, d_1, \cdots, d_N$ are

$$\begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_N \end{bmatrix} = \begin{bmatrix} -\hat{a} & 1 & \cdots & 0 \\ 0 & -\hat{a} & \ddots & \vdots \\ \vdots & \vdots & \ddots & N \\ 0 & 0 & \cdots & -\hat{a} \end{bmatrix}^{-1} \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_N \end{bmatrix},$$

and the optimal integration constant, also known as the initial condition, is obtained by least squares as

$$\hat{c} = \frac{\mathrm{e}^{2\hat{a}} - 1}{\mathrm{e}^{2\hat{a}n} - 1} \sum_{k=1}^{n} \left(x^{(1)}(k) - \sum_{j=0}^{N} d_j k^j \right) \mathrm{e}^{\hat{a}(k-2)}.$$

By substituting the time points into the time response function equation (5) and then using the inverse accumulating generation operator, the fitted and predicted values of the original series are expressed as

$$\hat{x}^{(0)}(k) = \begin{cases} \hat{x}^{(1)}(k) = \hat{c} + \sum\limits_{j=1}^{N} d_j, & k = 1, \\ \hat{x}^{(1)}(k) - \hat{x}^{(1)}(k-1), & k = 2, 3, \cdots. \end{cases} \tag{6}$$

It can be seen in the above modeling process that the time response function in (5) is obtained under the hypothesis that there exist no discretization error between the continuous equation (3) and the discrete equation (4) at time points. However, the differential equation (3) is discretized based on the intermediate value theorem (the trapezoid rule with $\lambda = 0.5$), which definitely introduces the discretization error (Wei et al., 2018). Therefore, the least square estimates in equation (4) can not be substituted into the time response function (5) unless the discretization error is quantified to guarantee the accuracy.

2.2. The representation of discrete grey polynomial model

We now turn to the difference equation (3) to simulate the accumulating generation series straightforward, the procedures are detailedly carried out as follows.

From the definition of accumulating generation series in equation (1), the first term of the right side in equation (3) can be expressed as

$$(1-\lambda)x^{(1)}(k-1) + \lambda x^{(1)}(k) = x^{(1)}(k-1) + \lambda x^{(0)}(k), \tag{7}$$

according to the binomial theorem,

$$k^{j+1} - (k-1)^{j+1} = \sum_{r=1}^{j+1} \binom{j+1}{r} (-1)^{r-1} k^{j-r+1}, \ j = 0, 1, \cdots, N, \tag{8}$$

then the second term of the right side in equation (3) is transformed into

$$\sum_{j=0}^{N} \frac{b_j}{j+1} \sum_{r=1}^{j+1} \binom{j+1}{r} (-1)^{r-1} k^{j-r+1} = \sum_{j=0}^{N} \left[\sum_{r=j}^{N} \frac{b_r}{r+1} \binom{r+1}{r-j+1} (-1)^{r-j} \right] k^j. \tag{9}$$

Substituting equations (7) and (9) into equation (3) gives that

$$x^{(0)}(k) = a x^{(1)}(k-1) + a\lambda x^{(0)}(k) + \sum_{j=0}^{N} \left[\sum_{r=j}^{N} \frac{b_r}{r+1} \binom{r+1}{r-j+1} (-1)^{r-j} \right] k^j + \varepsilon_k, \tag{10}$$

which is equivalent to

$$x^{(0)}(k) = \frac{a}{1-a\lambda} x^{(1)}(k-1) + \sum_{j=0}^{N} \left[\frac{1}{1-a\lambda} \sum_{r=j}^{N} \frac{b_r}{r+1} \binom{r+1}{r-j+1} (-1)^{r-j} \right] k^j + \frac{1}{1-a\lambda} \varepsilon_k. \tag{11}$$

By setting

$$\alpha = \frac{a}{1-a\lambda}, \ \beta_j = \frac{1}{1-a\lambda} \sum_{r=j}^{N} \frac{b_r}{r+1} \binom{r+1}{r-j+1} (-1)^{r-j}, \ \epsilon_k = \frac{1}{1-a\lambda} \varepsilon_k, \tag{12}$$

the discrete grey polynomial model, termed as DGPM(1,1,N), can be represented as

$$x^{(0)}(k) = \alpha x^{(1)}(k-1) + \beta_0 + \beta_1 k + \cdots + \beta_N k^N + \epsilon_k, \tag{13}$$

or, in a vector form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{14}$$

where

$$\boldsymbol{y} = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ x^{(0)}(4) \\ \vdots \\ x^{(0)}(n) \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} x^{(1)}(1) & 1 & 2 & \cdots & 2^N \\ x^{(1)}(2) & 1 & 3 & \cdots & 3^N \\ x^{(1)}(3) & 1 & 4 & \cdots & 4^N \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x^{(1)}(n-1) & 1 & n & \cdots & n^N \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{bmatrix}, \ \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

2.3. The parameters estimation of discrete grey polynomial model

It can be seen that equation (13) can be regarded as a special multiple regression model which combines the partial characteristics of autoregression and polynomial regression. Divide the matrix $\boldsymbol{X}$ into two parts, $\boldsymbol{X} = [\boldsymbol{u} \quad \boldsymbol{V}]$, where $\boldsymbol{u}$ is the first column and $\boldsymbol{V}$ is the rest $N+1$ independent columns. In general $\boldsymbol{u}$ is not in the column space of $\boldsymbol{V}$, and thus the columns of matrix $\boldsymbol{X}$ are linearly independent. Then, the least square estimates of model parameters can be obtained by minimizing the sum of squared error $\|\boldsymbol{\epsilon}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ and expressed as

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\alpha} & \hat{\beta}_0 & \hat{\beta}_1 & \cdots & \hat{\beta}_N \end{bmatrix}^{\mathrm{T}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}. \tag{15}$$

Substituting the least square estimates into the equation (14), the fitted values of the original time series excluding the first sample can be calculated as

$$\hat{X}^{(0)} = \hat{\boldsymbol{y}} = \left\{\hat{x}^{(0)}(2), \hat{x}^{(0)}(3), \cdots, \hat{x}^{(0)}(n)\right\} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \tag{16}$$

and the predicted values based on the recursive equation (13) can be expressed as

$$\hat{x}^{(0)}(n+\ell) = \begin{cases} \hat{\alpha}x^{(1)}(n) + \sum\limits_{j=0}^{N}\hat{\beta}_j(n+1)^j, & \ell = 1, \\ \hat{\alpha}\left[x^{(1)}(n) + \sum\limits_{i=1}^{\ell-1}\hat{x}^{(0)}(n+i)\right] + \sum\limits_{j=0}^{N}\hat{\beta}_j(n+\ell)^j, & \ell \geq 2. \end{cases} \tag{17}$$

2.4. Some comparisons with classical modeling approach

On the basis of the classical modeling approach in the conventional discrete grey models such as DGM(1,1) (Xie & Liu, 2009) and NDGM(1,1) (Xie et al., 2013a), we may generalize the equation forms of these two models and then deduce another similar expression of DGPM(1,1,N) model expressed as

$$x^{(1)}(k) = \varphi x^{(1)}(k-1) + \phi_0 + \phi_1 k + \cdots + \phi_N k^N + \varepsilon_k, \tag{18}$$

where the model parameters and the initial value are obtained by using a two-step method. In the first step, the model parameters are estimated by minimizing the sum of squared errors

$$\underset{\varphi,\phi_0,\cdots,\phi_N}{\arg\min} \sum_{k=2}^{n}\left[x^{(1)}(k) - \varphi x^{(1)}(k-1) - \sum_{j=0}^{N}\phi_j k^j\right]^2$$

where the initial value $\hat{x}^{(1)}(1) = x^{(1)}(1)$ is an invariant. In the second step, the initial value is assumed to be a variable and obtained by minimizing the sum of squared errors

$$\underset{\delta}{\arg\min} \sum_{k=2}^{n}\left[x^{(1)}(k) - \hat{\varphi}\hat{x}^{(1)}(k-1) - \sum_{j=0}^{N}\hat{\phi}_j k^j\right]^2$$

where $\hat{x}^{(1)}(1) = x^{(1)}(1) + \delta$ and δ is a variable. The inconsistency of the objective functions in these two steps may introduce additional errors.

It should be noticed that the proposed model and the two-step model are relevant by the following similar or different points:

(1) The accumulating generation is employed in two different models. The two-step model can be viewed as the process to find the relationship between the current cumulative sum and that at the last time points, and thus the inverse accumulating generation is essential to restore the fitting and predicting results of the accumulating generation series. However, the proposed model can be considered as the process to find the relationship between the current value and the cumulative sum at the last time point, so we can obtain the fitting and predicting results of the original series. It is obvious that the proposed model has a simpler modeling procedure without using the inverse accumulating generation.

(2) The model parameters in the two-step and the proposed models are respectively estimated by minimizing the objective functions expressed as

$$\|\boldsymbol{\epsilon}\|_2^2 = \sum_{k=2}^{n} \left[x^{(1)}(k) - \hat{x}^{(1)}(k)\right]^2 \quad \text{and} \quad \|\boldsymbol{\varepsilon}\|_2^2 = \sum_{k=2}^{n} \left[x^{(0)}(k) - \hat{x}^{(0)}(k)\right]^2.$$

Although these two objective functions are different, there exist fixed quantitative relationships between their estimated model parameters (see Theorem 1).

(3) The initial values in both models are selected by using two different approaches, thereby leading to the difference between their modeling results. In the two-step models, the fitting and predicting results are calculated based on the analytic solution and inverse accumulating generation, that is

$$\begin{cases} \hat{x}^{(1)}(1) = x^{(1)}(1) + \hat{\delta}, \\ \hat{x}^{(1)}(k) = \hat{\varphi}\hat{x}^{(1)}(k-1) + \sum\limits_{j=0}^{N} \hat{\phi}_j k^j, \\ \hat{x}^{(0)}(k) = \hat{x}^{(1)}(k) - \hat{x}^{(1)}(k-1), \ k = 2, 3, \cdots. \end{cases}$$

However, it can be seen in equations (16) and (17) that the proposed model uses a totally different approach which has already been validated and commonly used in time series analysis and forecasting (Box et al., 1994). The proposed model avoids the use of the inverse accumulating generation, and thus reduces unnecessary complexity for quantification of residual errors and interpretation of modeling results.

Theorem 1. *DGPM(1,1,N) model defined by equation (13) is equivalent to that defined by equation (18).*

Proof. Adding $x^{(1)}(k-1)$ on both sides of equation (13), then

$$x^{(0)}(k) + x^{(1)}(k-1) = x^{(1)}(k) = (\alpha+1)x^{(1)}(k-1) + \beta_0 + \beta_1 k + \cdots + \beta_N k^N + \epsilon_k$$

By setting $\varphi = \alpha + 1$, $\beta_j = \phi_j$, equation (18) is obtained and the least square estimates of its model parameters are expressed as

$$\hat{\boldsymbol{\phi}} = \begin{bmatrix} \varphi & \phi_0 & \phi_1 & \cdots & \phi_N \end{bmatrix}^{\mathrm{T}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{z}, \tag{19}$$

where

$$\boldsymbol{z} = \begin{bmatrix} x^{(1)}(2) & x^{(1)}(3) & x^{(1)}(4) & \cdots & x^{(1)}(n) \end{bmatrix}^{\mathrm{T}} = \boldsymbol{u} + \boldsymbol{y}. \tag{20}$$

Substituting equation (20) into the right side of equation (19) gives that

$$\hat{\boldsymbol{\phi}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u} + \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{e}_1 + \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \tag{21}$$

where $\boldsymbol{e}_1$ is a $(N+2) \times 1$ vector having 1 in the first entry and 0's in the other entries.

Substituting equation (15) into (21), the least square estimates of the model parameters in equations (13) and (18) satisfy

$$\hat{\boldsymbol{\phi}} = \boldsymbol{e}_1 + \hat{\boldsymbol{\beta}} \quad \Leftrightarrow \quad \hat{\varphi} = \hat{\alpha} + 1,\ \hat{\phi}_j = \hat{\beta}_j, \tag{22}$$

and the fitted and predicted values based on equations (15) and (16) satisfy

$$\hat{x}^{(1)}(k) = \hat{\varphi}\hat{x}^{(1)}(k-1) + \sum_{j=0}^{N} \hat{\phi}_j k^j \quad \Leftrightarrow \quad \hat{x}^{(0)}(k) = \hat{\alpha}\hat{x}^{(1)}(k-1) + \sum_{j=0}^{N} \hat{\beta}_j k^j. \tag{23}$$

By setting the initial value as that in equation (17), equations (13) and (18) are totally equivalent not only in the analytic expression but also from the application standpoint. Furthermore, when the order is $N = 0$, DGPM(1,1,N) yields to the DGM(1,1) model; when the order is $N = 1$, DGPM(1,1,N) yields to the NDGM(1,1) model.

Without regard to the initial value selection, the grey prediction models with discrete forms are equivalent to those with connotation forms, that is, GM(1,1,C) (Liu et al., 2014) and DGM(1,1) (Xie & Liu, 2009), SAIGM (Zeng et al., 2016) and NDGM(1,1) (Xie et al., 2013a), DGM(1,1,N) (Xie et al., 2013b) and FOTP-GM(1,1) (Li et al., 2018) are respectively equivalent to each other. □

3. Structure selection for discrete grey polynomial model

The modeling procedures in Section 2.3 are based on the hypothesis that the polynomial order is known in advance, which limits the application in practice. From both theoretical and practical perspectives, it is always not satisfactory on the basis of the following reasons:

(1) If the number of rows $n-1$ (n is the number of samples) is slightly greater than the number of columns $N+1$ (N is the polynomial order) in equation (14), there may be a lot of variability in the least square estimates in equation (15) (Hastie et al., 2013). (See example 1 in Section 4.3.)

(2) For higher polynomial order, it is likely to lead to the over-fitting problem. The fitted curve oscillates so wildly especially near the ends of original time series that we may end up with better fitting but poor predicting (Hastie et al., 2013). (See simulation studies in Section 5.2.)

In order to overcome these defects, the polynomial order is presupposed to be $N = 4$, and then the least square estimates in equation (15) are regained by minimizing the following objective function

$$\sum_{k=2}^{n} \left[x^{(0)}(k) - \alpha x^{(1)}(k-1) - \beta_0 - \sum_{j=1}^{N} \beta_j k^j \right]^2 \quad \text{subject to} \quad \sum_{j=1}^{N} \mathcal{I}(\beta_j \neq 0) = s \tag{24}$$

where $0 \leq s \leq N$ is a tuned factor controlling the model structure; $\mathcal{I}(\cdot)$ is an indicator function that returns a 1 if the condition is true, and returns a 0 otherwise.

In fact, the nature of minimizing the objective function under the constraint condition in equation (24) can be considered as a best subset selection context, and the tuned factor $s \in \{0, 1, 2, 3, 4\}$ controls the subset size. That is to say, all the $2^N = 16$ possible models are fitted and evaluated separately. Then the validation set approach is employed to estimate the predicting error of each model, and the original series is split into two parts: a training set used to build model and a validation set used to be predicted. Taking the mean absolute percentage error (*mape*) as the performance measure, both the fitting and predicting errors can be obtained from

$$mape = \frac{1}{n_2 - n_1 + 1} \sum_{k=n_1}^{n_2} \left| \frac{x^{(0)}(k) - \hat{x}^{(0)}(k)}{x^{(0)}(k)} \right| \times 100\%, \tag{25}$$

where $n_1 < n_2$ and $n_1, n_2 \in \mathbb{N}$. If $n_1 = 2$ and $n_2 = m$, then $mape = mape_{\text{train}}$; if $n_1 = m + 1$ and $n_2 = n$, then $mape = mape_{\text{valid}}$.

In detail, the best subset selection for selecting the optimal model from among all the possibilities is described in Algorithm 1.

Algorithm 1 Data-based structure selection (DBSS) algorithm

Input: Original time series $X^{(0)}$, polynomial order N.

1: Divide the original series into two parts:

$$X^{(0)}_{\text{train}} = \left\{x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(m)\right\} \text{ and } X^{(0)}_{\text{valid}} = \left\{x^{(0)}(m+1), x^{(0)}(m+2), \cdots, x^{(0)}(n)\right\};$$

where m is the maximal integer that is less than or equal to $\frac{4}{5}n$ (Hastie et al., 2013).

2: Let $\mathcal{M}_{0,1}$ denote the null model ($s = 0$) that only contains the accumulative and constant terms in the right side of equation (13).

3: **for all** $s = 1, 2, \cdots, N$ **do**

4: Fit all $\binom{N}{s}$ models that contain exactly $s+2$ terms (1 accumulative term, 1 constant term and s other terms) in the right side of equation (13).

5: Let $\mathcal{M}_{s,t}$ denote the superior picked from the above $\binom{N}{s}$ models. Here superior is defined as

$$\mathcal{M}_{s,q} = \arg\min_{\mathcal{M}} \{\mathcal{M}(mape_{\text{train}}) \leq \eta_1\}, \ q = 1, 2, \cdots$$

where $\eta_1 \in (0, 1]$ controls the fitting error and traditionally set as 0.1 in application.

6: **end for**

7: Pick a single optimal model from among $\mathcal{M}_{s,q}$, and denote it as $\mathcal{M}_b$. Here optimum is defined as

$$\mathcal{M}_b = \arg\min_{\mathcal{M}} P(\mathcal{M}) \quad \text{subject to} \quad \mathcal{M}(mape_{\text{valid}}) \leq \eta_2,$$

where $P(\mathcal{M})$ denotes the order of polynomial in $\mathcal{M}$; $\eta_2 \in (0, 1]$ controls the predicting error and traditionally set as 0.1 in application.

Output: $\mathcal{M}_b$.

For the convenience of property analysis in the following, the optimal model $\mathcal{M}_b$ is expressed as the selective matrix. For example, if the optimal model is $x^{(0)}(k) = \alpha x^{(1)}(k-1) + \beta_0 + \beta_1 k$, the least square estimates are rewritten as

$$\hat{\boldsymbol{\beta}}_s = [\hat{\alpha} \quad \hat{\beta}_0]^{\mathrm{T}} = \left(\boldsymbol{X}_s^{\mathrm{T}} \boldsymbol{X}_s\right)^{-1} \boldsymbol{X}_s^{\mathrm{T}} \boldsymbol{y},$$

where $\boldsymbol{X}_s = \boldsymbol{X}\boldsymbol{S}$, $\boldsymbol{S} = [\boldsymbol{e}_1 \quad \boldsymbol{e}_2 \quad \boldsymbol{e}_3]$ is a $(N+2) \times 3$ selective matrix, and $\boldsymbol{e}_\iota$ is a $(N+2) \times 1$ vector having 1 in the ι-th entry and 0's in the other entries.

In the end, the optimal model $\mathcal{M}_b$ is fitted based on the original series ($X^{(0)}_{\text{train}} \bigcup X^{(0)}_{\text{valid}} = X^{(0)}$) to take full advantage of the samples, and the fitted values according to equation (16) are

$$\hat{X}^{(0)} = \hat{\boldsymbol{y}} = \left\{\hat{x}^{(0)}(2), \hat{x}^{(0)}(3), \cdots, \hat{x}^{(0)}(n)\right\} = \boldsymbol{X}\boldsymbol{S}_b\hat{\boldsymbol{\beta}}_b, \tag{26}$$

and the predicted values according to equation (17) are

$$\hat{x}^{(0)}(n+\ell) = \begin{bmatrix}\chi & 1 & (n+\ell) & \cdots & (n+\ell)^N\end{bmatrix} \boldsymbol{S}_b\hat{\boldsymbol{\beta}}_b, \tag{27}$$

where

$$\chi = \begin{cases} x^{(1)}(n), & \ell = 1, \\ x^{(1)}(n) + \sum_{i=1}^{\ell-1} \hat{x}^{(0)}(n+i), & \ell \geq 2. \end{cases}$$

To sum up, once the original series and the predicted length are given, the fitted and predicted values can be obtained according to equations (26) and (27).

4. Theoretical property analysis

In this section, the propositions are only proved in the case that the selective matrix $\boldsymbol{S}_b$ is equal to the identity matrix $\boldsymbol{I}$, and it is easy to prove that the conclusions still hold true for other selective matrices.

Lemma 1. *Assuming that* $X_a^{(1)} = \left\{x_a^{(1)}(1), x_a^{(1)}(2), \cdots, x_a^{(1)}(n)\right\}$, *where* $x_a^{(1)}(k) = \rho x^{(1)}(k) + \xi$, $\rho \neq 0$, *is the affine transformation of the accumulating generation series* $X^{(1)}$, *then the fitted and predicted values corresponding to* $X_a^{(1)}$ *and* $X^{(1)}$ *satisfy* $\hat{x}_a^{(0)}(k) = \rho\hat{x}^{(0)}(k)$, $k = 2, 3, \cdots$.

Proof. Combining the definitions of affine transformation and accumulation generation, it follows that

$$x_a^{(0)}(k) = x_a^{(1)}(k) - x_a^{(1)}(k-1) = \rho x^{(0)}(k),\ k = 2, 3, \cdots, n.$$

The vector $\boldsymbol{y}_a$ is obtained as

$$\boldsymbol{y}_a = \left[x_a^{(0)}(2),\ x_a^{(0)}(3),\ x_a^{(0)}(4),\ \cdots,\ x_a^{(0)}(n)\right]^{\mathrm{T}} = \rho\boldsymbol{y},$$

and the matrix $\boldsymbol{X}_a$ is decomposed into three parts

$$\boldsymbol{X}_a = \begin{bmatrix} x_a^{(1)}(1) & 1 & 2 & \cdots & 2^N \\ x_a^{(1)}(2) & 1 & 3 & \cdots & 3^N \\ x_a^{(1)}(3) & 1 & 4 & \cdots & 4^N \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_a^{(1)}(n-1) & 1 & n & \cdots & n^N \end{bmatrix} = \begin{bmatrix} \rho x^{(1)}(1) + \xi & 1 & 2 & \cdots & 2^N \\ \rho x^{(1)}(2) + \xi & 1 & 3 & \cdots & 3^N \\ \rho x^{(1)}(3) + \xi & 1 & 4 & \cdots & 4^N \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho x^{(1)}(n-1) + \xi & 1 & n & \cdots & n^N \end{bmatrix} = \boldsymbol{X}\boldsymbol{P}\boldsymbol{Q},$$

where $\boldsymbol{P}$ and $\boldsymbol{Q}$ are both $(N+2)$-order non-singular matrices respectively defined as

$$\boldsymbol{P} = \begin{bmatrix} \rho & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{Q} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \xi & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Let $\boldsymbol{\beta}_a = [\alpha' \quad \beta_0' \quad \beta_1' \quad \cdots \quad \beta_N']^{\mathrm{T}}$ be the model parameters of the DGPM(1,1,N) model corresponding to the affine transformation series $X_a^{(1)}$, then the least square estimates $\hat{\boldsymbol{\beta}}_a = [\hat{\alpha}' \quad \hat{\beta}_0' \quad \hat{\beta}_1' \quad \cdots \quad \hat{\beta}_N']^{\mathrm{T}}$ can be expressed as

$$\hat{\boldsymbol{\beta}}_a = \left(\boldsymbol{X}_a^{\mathrm{T}}\boldsymbol{X}_a\right)^{-1}\boldsymbol{X}_a^{\mathrm{T}}\boldsymbol{y}_a = \rho\boldsymbol{Q}^{-1}\boldsymbol{P}^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \rho\boldsymbol{Q}^{-1}\boldsymbol{P}^{-1}\hat{\boldsymbol{\beta}},$$

that is,

$$\hat{\alpha}' = \hat{\alpha},\ \hat{\beta}_0' = -\hat{\alpha}\xi + \rho\hat{\beta}_0,\ \hat{\beta}_j' = \rho\hat{\beta}_j,\ j = 1, 2, \cdots, N. \tag{28}$$

Substituting the least square estimates in equation (28) into the fitted and predicted values based on equations (26) and (27) gives that

$$\begin{aligned} \hat{X}_a^{(0)} = \hat{\boldsymbol{y}}_a &= \left\{\hat{x}_a^{(0)}(2), \hat{x}_a^{(0)}(3), \cdots, \hat{x}_a^{(0)}(n)\right\} = \boldsymbol{X}_a\hat{\boldsymbol{\beta}}_a = (\boldsymbol{X}\boldsymbol{P}\boldsymbol{Q})\left(\rho\boldsymbol{Q}^{-1}\boldsymbol{P}^{-1}\hat{\boldsymbol{\beta}}\right) = \rho\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \rho\hat{\boldsymbol{y}} = \left\{\rho\hat{x}^{(0)}(2), \rho\hat{x}^{(0)}(3), \cdots, \rho\hat{x}^{(0)}(n)\right\}, \end{aligned}$$

and

$$\hat{x}_a^{(0)}(n+\ell) = \begin{cases} \hat{\alpha}' x_a^{(1)}(n) + \sum\limits_{j=0}^{N}\hat{\beta}_j'(n+1)^j = \rho\hat{x}^{(0)}(n+1), & \ell = 1, \\ \hat{\alpha}'\left[x_a^{(1)}(n) + \sum\limits_{i=1}^{\ell-1}\hat{x}_a^{(0)}(n+i)\right] + \sum\limits_{j=0}^{N}\hat{\beta}_j'(n+\ell)^j = \rho\hat{x}^{(0)}(n+\ell), & \ell \geq 2. \end{cases}$$

Lemma 1 shows that all the data normalization methods that can be expressed as the affine transformations have no influence on the modeling performance, for example, the feature scaling method in which the multiple coefficient $\rho = 1/(max - min)$ and the translation coefficient $\xi = -min/(max - min)$, where $max = \max_{k=1}^{n} \left\{x^{(1)}(k)\right\}$, $min = \min_{k=1}^{n} \left\{x^{(1)}(k)\right\}$. □

4.1. The influence of multiple transformation of original series on modeling

Theorem 2. *Assuming that* $X_m^{(0)} = \left\{x_m^{(0)}(1), x_m^{(0)}(2), \cdots, x_m^{(0)}(n)\right\}$, *where* $x_m^{(0)}(k) = \rho x^{(0)}(k)$, $k = 1, 2, \cdots, n$, *is the multiple transformation of original series* $X^{(0)}$, *then the fitted and predicted values corresponding to* $X_m^{(0)}$ *and* $X^{(0)}$ *satisfy* $\hat{x}_m^{(0)}(k) = \rho\hat{x}^{(0)}(k)$, $k = 2, 3, \cdots$.

Proof. According to equation (1), the accumulating generation of the multiple transformation series is calculated as

$$x_m^{(1)}(k) = \sum_{i=1}^{k} x_m^{(0)}(i) = \rho \sum_{i=1}^{k} x^{(0)}(i) = \rho x^{(1)}(k), \ k = 1, 2, \cdots, n,$$

which can be viewed as a special form of $x_a^{(1)}(k)$ in Lemma 1 (the translation coefficient $\xi = 0$). Therefore, the fitted and predicted values satisfy $\hat{x}_m^{(0)}(k) = \rho\hat{x}^{(0)}(k) + \xi = \rho\hat{x}^{(0)}(k)$.

In addition, if the first sample satisfies $x_m^{(0)}(1) = \rho x^{(0)}(1) + \xi$ and the rests are $x_m^{(0)}(k) = \rho x^{(0)}(k)$, then

$$x_m^{(1)}(k) = \sum_{i=1}^{k} x_m^{(0)}(i) = \rho \sum_{i=1}^{k} x^{(0)}(i) + \xi = \rho x^{(1)}(k) + \xi, \ k = 1, 2, \cdots, n,$$

which is equivalent to the affine transformation $x_a^{(1)}(k)$ in Lemma 1, and thus the conclusion hold true.

Theorems 2 shows that the values of the first sample in the original series and the multiple coefficient in multiple transformation have no influence on the fitted and predicted values. Hence, we can select a suitable multiple coefficient ρ to reduce the condition number in real world applications. □

4.2. Unbias for original series composed of pure exponential and polynomial functions

Theorem 3. *Assuming that* $X^{(0)} = \left\{x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)\right\}$, *where* $x^{(0)}(k) = ce^{ak} + b_0 + b_1 k + \cdots + b_N k^N$, $c \neq 0$, $k = 1, 2, \cdots, n$, *is the original series, then the fitted and predicted values by DGPM(1,1,N+1) model satisfy* $\hat{x}^{(0)}(k) = x^{(0)}(k)$, $k = 2, 3, \cdots$.

Proof. Let $\mu = e^a$, $\nu_j = \frac{b_j}{c}$, then we only need to prove the conclusion to be true with $x^{(0)}(k) = \mu^k + \nu_0 + \nu_1 k + \cdots + \nu_N k^N$.

Without loss of generality, let $N = 3$, then according to equation (15) the parameter estimates corresponding to DGPM(1,1,4) model are obtained as

$$\hat{\alpha} = \mu - 1, \ \hat{\beta}_0 = \mu(\nu_0 + 1), \ \hat{\beta}_1 = \nu_0(1 - \mu) + \frac{1}{2}\nu_1(1 + \mu) + \frac{1}{6}\nu_2(1 - \mu),$$

$$\hat{\beta}_2 = \frac{1}{2}\nu_1(1 - \mu) + \frac{1}{2}\nu_2(1 + \mu) + \frac{1}{4}\nu_3(1 - \mu), \ \hat{\beta}_3 = \frac{1}{3}\nu_2(1 - \mu) + \frac{1}{2}\nu_3(1 + \mu), \ \hat{\beta}_4 = \frac{1}{4}\nu_3(1 - \mu).$$

Substituting parameter estimates into equations (15) and (16) gives the fitted values expressed as

$$\hat{X}^{(0)} = \hat{\boldsymbol{y}} = \left\{\hat{x}^{(0)}(2), \hat{x}^{(0)}(3), \cdots, \hat{x}^{(0)}(n)\right\} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} = \left\{x^{(0)}(2), x^{(0)}(3), \cdots, x^{(0)}(n)\right\},$$

and the predicted values expressed as

$$\hat{x}^{(0)}(n + \ell) = \begin{cases} \hat{\alpha}x^{(1)}(n) + \sum\limits_{j=0}^{N} \hat{\beta}_j (n + 1)^j = x^{(0)}(n + 1), & \ell = 1, \\ \hat{\alpha}\left[x^{(1)}(n) + \sum\limits_{i=1}^{\ell-1} \hat{x}^{(0)}(n + i)\right] + \sum\limits_{j=0}^{N} \hat{\beta}_j (n + \ell)^j = x^{(0)}(n + \ell), & \ell \geq 2. \end{cases}$$

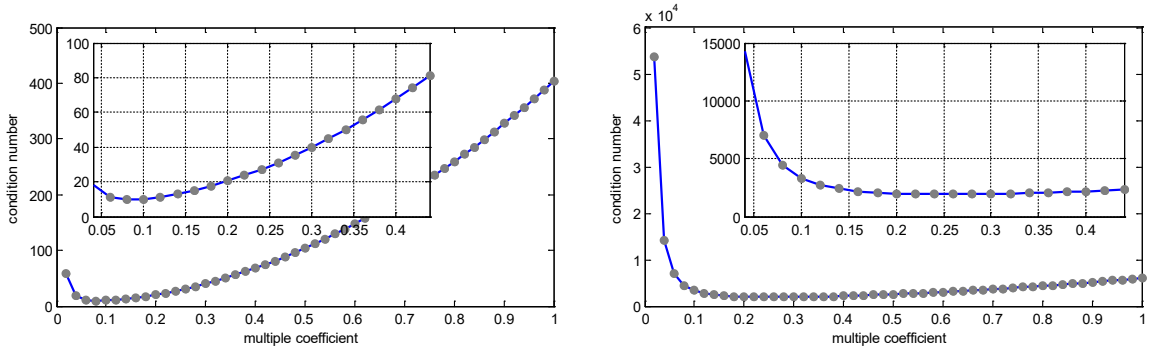
In particular, if $\mu = 1$, then $\hat{\alpha} = 0$, $\hat{\beta}_0 = \nu_0 + 1$, $\hat{\beta}_1 = \nu_1$, $\hat{\beta}_2 = \nu_2$, $\hat{\beta}_3 = \nu_3$, $\hat{\beta}_4 = \nu_4$. DGPM(1,1,4) yields to the fourth degree polynomial regression model, and the conclusion still holds true.

Similarly, if $\nu_3 = 0$, then $\hat{\alpha} = \mu - 1$, $\hat{\beta}_0 = \mu(\nu_0 + 1)$, $\hat{\beta}_1 = \nu_0(1-\mu) + \frac{1}{2}\nu_1(1+\mu) + \frac{1}{6}\nu_2(1-\mu)$, $\hat{\beta}_2 = \frac{1}{2}\nu_1(1-\mu) + \frac{1}{2}\nu_2(1+\mu)$, $\hat{\beta}_3 = \frac{1}{3}\nu_2(1-\mu)$. DGPM(1,1,4) yields to the DGPM(1,1,3) model, and the conclusion still holds true. That is, DGPM(1,1,3) is unbiased for the series $X^{(0)} = \left\{x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)\right\}$, where $x^{(0)}(k) = ce^{ak} + b_0 + b_1 k + b_2 k^2$.

Theorem 3 shows that the fitted and predicted values by DGPM(1,1,N) model are unbiased when the original series follows the distribution with partial exponential and partial polynomial characteristics. □

4.3. Numerical examples for property validation

Example 1. *Assuming that the original time series is* $X^{(0)} = \{1.2,\ 1.4,\ 1.8,\ 2.3,\ 2.7,\ 3.3,\ 4.2,\ 4.8,\ 6.1\}$, *the multiple coefficient is taken at every 0.02 units on the interval* $(0, 1.0]$. *Figure 1 displays the condition number of* $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ *when estimating the model parameters of DGPM(1,1,0) and DGPM(1,1,1) model. The results shows that there indeed exists a multiple coefficient to minimize the condition number in this numerical example. Moreover, although the error of DGPM(1,1,0) model (mape* $= 2.32\%$*) is a little less than that of DGPM(1,1,1) model (mape* $= 2.47\%$*), the condition numbers of the former in Figure 1(a) are much less than those of the latter in Figure 1(b) under every multiple coefficient, which indicates that DGPM(1,1,1) is a ill-conditioned model, and also illustrates the importance and necessity of model selection. The condition numbers are all computed with respect to the* L^1 *matrix norm (maximum absolute column sum).*



(a) DGPM(1,1,0) model (b) DGPM(1,1,1) model

Figure 1: Trends of the condition number with the increasing multiple coefficient in DGPM(1,1,N) models with the order $N = 0, 1$.

Example 2. *From Zeng (2018),* $X^{(0)} = \{17.72,\ 27.39,\ 45.09,\ 84.60,\ 183.41,\ 443.43,\ 1141.63\}$ *is sampled from* $x^{(0)}(k) = e^k + 5k + 10$ *at the time points* $k = 1, 2, \cdots, 7$. *Then form Theorem 3, the polynomial order is set as* $N = 2$, *that is, DGPM(1,1,2) model are constructed based on the first 6 samples. The corresponding estimates of model parameters from equation (15) are obtained as*

$$\hat{\alpha} = e - 1 = 1.718282,\ \ \hat{\beta}_0 = 11e = 29.901100,$$
$$\hat{\beta}_1 = 12.5 - 7.5e = -7.887114,\ \ \hat{\beta}_2 = 2.5(1-e) = -4.295705,$$

and the fitted and predicted values form equations (16) and (17) are calculated as

$$\left\{\hat{x}^{(0)}(2), \hat{x}^{(0)}(3), \hat{x}^{(0)}(4), \hat{x}^{(0)}(5), \hat{x}^{(0)}(6), \hat{x}^{(0)}(7)\right\} = \{27.39, 45.09, 84.60, 183.41, 443.43, 1141.63\}.$$

The above results not only validate the conclusion in Theorem (3) but also reveal that DGPM(1,1,N*) is more accurate than the grey model with fractional order accumulation in Zeng (2018).*

5. Simulations

5.1. Data generation mechanism

In order to further validate the theoretical analysis, we design a simulation study where the original time series is generated by adding random error to a mixture of exponential and linear functions, that is

$$x^{(0)}(t) = e^{at} + b_0 + b_1 t + \varepsilon_t,$$

where $\varepsilon_t \backsim \mathcal{N}(0, \sigma^2)$ is the random error.

The model parameters are set as $(a, b_0, b_1) = (0.5, 5.0, -2.0)$ to generate the samples at every time interval of 0.25 in the range of $t \in [0, 7]$, and the random errors are generated based on the normal distribution with mean 0 and standard deviation varying $\sigma = 0.05$, 0.06, 0.07, 0.08, 0.09, 0.10. Then DGPM(1,1,N) models with polynomial order varying $N = 0$, 1, 2, 3 are respectively constructed, and therefore we have a total of 24 scenarios of different combinations.

5.2. Modeling results and comparisons

In each scenario, the original series is split into two parts: the in-sample period (the first 23 samples) used to fit models and the out-of-sample period (the rest 6 samples) used to assess the predicting error, and 500 runs are replicated using the statistical computing software **R** (version 3.4.3). The total averages of the mean absolute percentage errors in the in-sample (MAPE_{in}) and out-of-sample (MAPE_{out}) periods are used to compare the performance of different models:

$$\text{MAPE}_{\text{in}} = \frac{1}{500}\sum_{i=1}^{500} mape_{\text{in}} \quad \text{and} \quad \text{MAPE}_{\text{out}} = \frac{1}{500}\sum_{i=1}^{500} mape_{\text{out}},$$

where

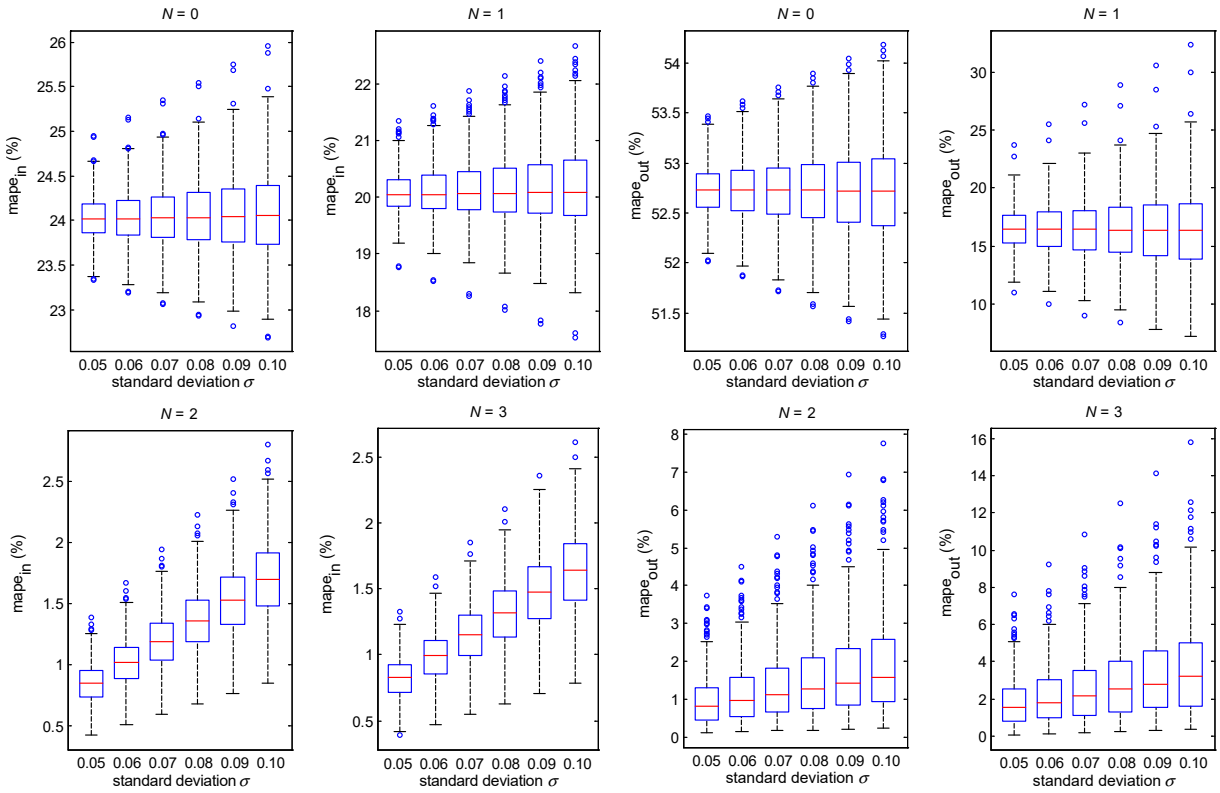
$$mape_{\text{in}} = \frac{1}{22}\sum_{k=2}^{23}\left|\frac{x^{(0)}(k) - \hat{x}^{(0)}(k)}{x^{(0)}(k)}\right| \times 100\% \quad \text{and} \quad mape_{\text{out}} = \frac{1}{6}\sum_{k=24}^{29}\left|\frac{x^{(0)}(k) - \hat{x}^{(0)}(k)}{x^{(0)}(k)}\right| \times 100\%.$$

Figure 2 shows the boxplots of the fitting and predicting errors in the 500 simulation replications, and also the averages of the errors in Table 1.

Table 1: Comparison of the four DGPM(1,1,N) models with the order $N = 0, 1, 2, 3$ in terms of the in-sample error MAPE_{in} and the out-of-sample error MAPE_{out} with the standard deviation of random error $\sigma = 0.05, 0.06, 0.07, 0.08, 0.09, 0.10$.

Errors	Order	Standard deviation					
		0.05	0.06	0.07	0.08	0.09	0.10
MAPE_{in}	0	24.02	24.03	24.04	24.05	24.06	24.07
	1	20.08	20.09	20.11	20.12	20.14	20.15
	2	0.85	1.02	1.19	1.35	1.52	1.69
	3	0.82	0.99	1.15	1.31	1.48	1.64
MAPE_{out}	0	52.73	52.73	52.72	52.72	52.71	52.71
	1	16.53	16.54	16.55	16.55	16.55	16.55
	2	0.96	1.15	1.34	1.53	1.72	1.92
	3	1.79	2.15	2.51	2.87	3.22	3.59

Figure 1 and Table 1 show that for the scenarios with the same polynomial order, both the fitting error and the predicting error tend to increase with the increase of the standard deviation of random error; for the scenarios with the same standard deviation of random error, both the fitting error and the predicting error decrease firstly to the minimum and then increase with the increase of the polynomial order; for all the scenarios, DGPM(1,1,2) and DGPM(1,1,3) perform far better than the rest two models in terms of both the fitting error and the predicting error. Furthermore, DGPM(1,1,2) and DGPM(1,1,3) perform almost the



(a) Fitting error (b) Predicting error

Figure 2: Boxplots for the fitting error $mape_{\text{in}}$ and the predicting error $mape_{\text{out}}$ by using the four DGPM(1,1,N) models with the order $N = 0, 1, 2, 3$ in 500 simulation replications with the standard deviation of random error $\sigma = 0.05, 0.06, 0.07, 0.08, 0.09, 0.10$.

same especially in terms of the fitting performance, that is, the difference of MAPE_{out}s is less than 2.00% and that of MAPE_{in}s is even less than 0.05%.

Because the $mape_{\text{in}}$s and $mape_{\text{out}}$s of DGPM(1,1,2) and DGPM(1,1,3) are respectively paired (two models on each individual time series), we consider the paired t-test for a parametric test. Table 2 shows that there exist significant differences between these two models. From a statistical perspective, the mean of $mape_{\text{in}}$s of DGPM(1,1,2) is significantly greater than that of DGPM(1,1,3) in the in-sample period, whereas the mean of $mape_{\text{out}}$s of DGPM(1,1,2) is significantly less than that of DGPM(1,1,3) in the out-of-sample period.

Table 2: Significance of comparisons of the second-order and third-order models with the number of simulation replications 25, 50, 100, 500, and the standard deviation of random error $\sigma = 0.05, 0.06, 0.07, 0.08, 0.09, 0.10$.

Alternative hypothesis	Replications	Standard deviation					
		0.05	0.06	0.07	0.08	0.09	0.10
$\mu_2 - \mu_3 > 0$ [a]	25	0.02 (0.003*)	0.03 (0.003*)	0.03 (0.003*)	0.04 (0.003*)	0.05 (0.002*)	0.06 (0.002*)
	50	0.02 (0.000*)	0.03 (0.000*)	0.04 (0.000*)	0.05 (0.000*)	0.05 (0.000*)	0.06 (0.000*)
	100	0.03 (0.000*)	0.03 (0.000*)	0.04 (0.000*)	0.05 (0.000*)	0.05 (0.000*)	0.06 (0.000*)
	500	0.02 (0.000*)	0.03 (0.000*)	0.03 (0.000*)	0.04 (0.000*)	0.05 (0.000*)	0.05 (0.000*)
$\mu_2' - \mu_3' < 0$ [b]	25	−0.63 (0.004*)	−0.77 (0.003*)	−0.91 (0.002*)	−1.06 (0.002*)	−1.23 (0.001*)	−1.40 (0.001*)
	50	−0.80 (0.000*)	−0.97 (0.000*)	−1.13 (0.000*)	−1.31 (0.000*)	−1.48 (0.000*)	−1.67 (0.000*)
	100	−0.64 (0.000*)	−0.77 (0.000*)	−0.91 (0.000*)	−1.04 (0.000*)	−1.18 (0.000*)	−1.32 (0.000*)
	500	−0.83 (0.000*)	−1.00 (0.000*)	−1.16 (0.000*)	−1.33 (0.000*)	−1.50 (0.000*)	−1.67 (0.000*)

[a] The mean of $mape_{\text{in}}$s of DGPM(1,1,2) is greater than that of DGPM(1,1,3).
[b] The mean of $mape_{\text{out}}$s of DGPM(1,1,2) is less than that of DGPM(1,1,3).
[*] Highly significant at the 0.01 level (one-sided).

5.3. Short discussion

It can be seen from the data generation mechanism that the original time series is a mixture of exponential component, linear component and random error, and this pattern is in agreement with the unbiased property in Theorem 3 that the second-order model is unbiased for the time series composed of pure exponential and linear functions. Therefore, DGPM(1,1,2) should be the best even though there exists noise in the generated time series. The above simulation results show that DGPM(1,1,0) and DGPM(1,1,1) are both under-fitting models with higher fitting errors and predicting errors, and at the same time DGPM(1,1,3) is a over-fitting model with lower fitting error but higher predicting error. Only DGPM(1,1,2) performs best among these models, which also validates the robustness of the proposed model.

6. Real data tests

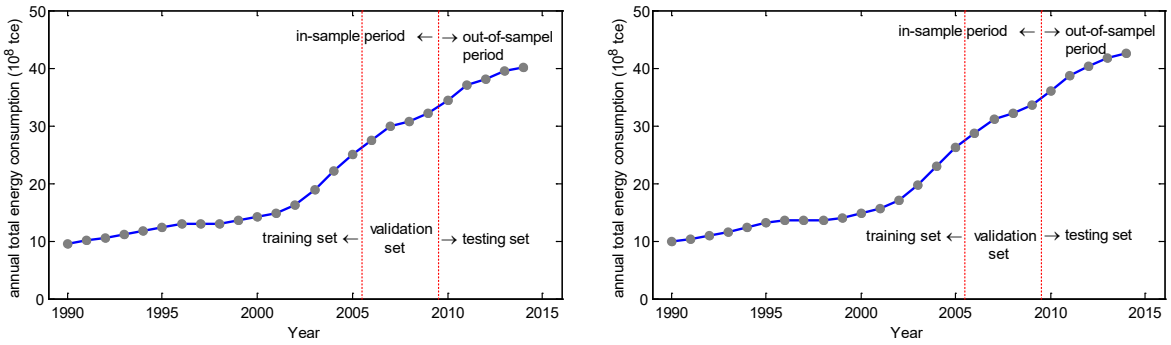
6.1. Data collection

Energy demand is the prerequisite condition for a nation's economic development, and it is significant to predict accurate energy consumption. There exist two accounting approaches for annual total energy consumption: calorific value calculation (ATEC_{cvc}) and coal equivalent calculation (ATEC_{cec}) in China.

The ATEC_{cvc} and ATEC_{cec} data of China in the period from 1990 to 2014 are respectively collected in Tables 3 and 4 (China Energy Statistical Yearbook 2015). In order to validate the performance of proposed model, the original series is divided into three parts: a training set used to fit the models, a validation set used to select the optimal model structure and a test set used for assessment of the predicting error of the final chosen model, as shown in Figures 3(a) and 3(b).

6.2. Modeling procedures and results

It can be seen in Tables 3 and 4 that although the values of ATEC_{cvc} and ATEC_{cec} are different at all time points, their trends (see Figure 3) and modeling procedures are similar with each other. Therefore,

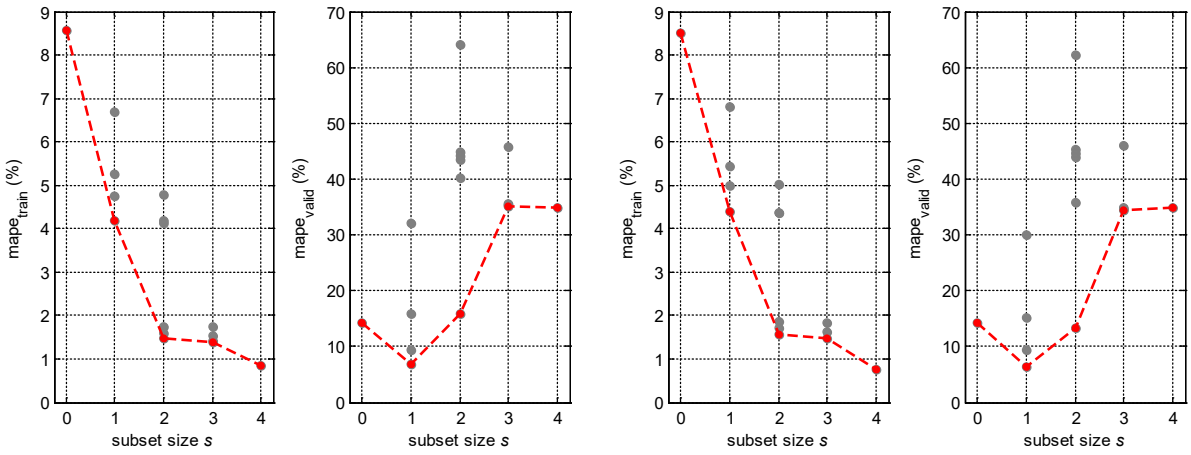


(a) ATEC_{cvc} (b) ATEC_{cec}

Figure 3: Split of the annual total energy consumption data (based on the calorific value calculation and coal equivalent calculation approaches respectively) in China from 1990 to 2014.

only the modeling details of ATEC_{cvc} are presented, and the results corresponding to ATEC_{cvc} are just briefly summarized in the following.

According to Algorithm 1 (where the polynomial order is set as $N = 4$), the models are separately fitted based on the training set from 1990 to 2005, the optimal model is selected based on the validation set from 2006 to 2009. The $mape_{\text{train}}$s and $mape_{\text{valid}}$s for each possible model are plotted in Figure 4.



(a) ATEC_{cvc} (b) ATEC_{cec}

Figure 4: Comparison of all possible subset models in terms of the training error $mape_{\text{train}}$ and the validation error $mape_{\text{valid}}$ in the data sets ATEC_{cvc} and ATEC_{cec}. The grey spot indicates the value of error for each possible model and the red frontier tracks the best model for a given number of subset size.

It can be seen in Figure 4(a) that as the subset size increases, the $mape_{\text{train}}$ decreases quickly from 8.6% to 0.9%, while the $mape_{\text{valid}}$ decreases firstly and then increases. The $mape_{\text{train}}$s of every possible model are less than 0.1, but only the $mape_{\text{valid}}$s of $\mathcal{M}_{1,2}$ with form $x^{(0)}(k) = \alpha x^{(1)}(k) + \beta_0 + \beta_2 k^2$ and $\mathcal{M}_{1,3}$ with form $x^{(0)}(k) = \alpha x^{(1)}(k) + \beta_0 + \beta_2 k^3$ are less than 0.1. The best model is selected as $\mathcal{M}_b = \mathcal{M}_{1,2}$ due to that the orders of polynomial satisfy $P(\mathcal{M}_{1,2}) = 2 < P(\mathcal{M}_{1,3}) = 3$. Fitting this optimal model on the series $X^{(0)}_{\text{train}} \bigcup X^{(0)}_{\text{valid}}$ gives the final model expressed as

$$\text{ATEC}_{\text{cvc}}: \ x^{(0)}(k) = -0.08569752x^{(1)}(k-1) + 11.07645 + 0.1206434k^2,$$

and also the fitted and predicted values computed by equations (26) and (27) listed in Table 3.

Similarly, the final model for ATEC_{cec} is obtained as

$$\text{ATEC}_{\text{cec}}: \ x^{(0)}(k) = -0.08146799x^{(1)}(k-1) + 11.42946 + 0.1227668k^2.$$

and also the fitted and predicted value displayed in Table 4.

Table 3: Comparison of the fitted results (from 1990 to 2009) and predicted results (from 2010 to 2014) by using various grey and other models in the data set ATEC_{cvc}.

Year	Actual	DGPM		NDGM(1,1)		NBGM(1,1) [a]		GPMB(1,1,0) [b]		QPR [c]		SVR [d]		NNAR [e]	
	values	Values	APE	Values	APE	Values	APE	Values	APE	Values	APE	Values	APE	Values	APE
1990	9.5384	—	—	9.9556	4.37	—	—	10.0743	5.62	10.8917	14.19	—	—	—	—
1991	10.0413	10.7416	6.97	9.8555	1.85	8.0124	20.21	8.2364	17.98	10.5928	5.49	—	—	—	—
1992	10.5602	10.4843	0.72	10.2077	3.34	8.6777	17.83	8.8952	15.77	10.4590	0.96	11.5210	9.10	10.7600	1.89
1993	11.1490	10.4238	6.50	10.6096	4.84	9.3850	15.82	9.6067	13.83	10.4905	5.91	11.5077	3.22	11.1262	0.20
1994	11.8071	10.5542	10.61	11.0682	6.26	10.1426	14.10	10.3751	12.13	10.6872	9.48	11.6667	1.19	11.5969	1.78
1995	12.3471	10.8694	11.97	11.5915	6.12	10.9565	11.26	11.2049	9.25	11.0491	10.51	12.0548	2.37	12.1965	1.22
1996	12.9665	11.3797	12.24	12.1886	6.00	11.8321	8.75	12.1012	6.67	11.5762	10.72	12.6175	2.69	12.7044	2.02
1997	13.0082	12.0781	7.15	12.8700	1.06	12.7750	1.79	13.0691	0.47	12.2686	5.69	13.3341	2.50	13.4233	3.19
1998	13.0260	13.0143	0.09	13.6475	4.77	13.7908	5.87	14.1144	8.36	13.1261	0.77	13.8049	5.98	13.3079	2.16
1999	13.5132	14.1902	5.01	14.5347	7.56	14.8854	10.15	15.2434	12.80	14.1489	4.70	13.8500	2.49	13.3234	1.40
2000	14.0993	15.5657	10.40	15.5471	10.27	16.0653	13.94	16.4626	16.76	15.3369	8.78	14.2320	0.94	14.0720	0.19
2001	14.8264	17.1322	15.55	16.7023	12.65	17.3373	16.94	17.7794	19.92	16.6901	12.57	15.1149	1.95	14.9458	0.81
2002	16.1935	18.8777	16.58	18.0205	11.28	18.7088	15.53	19.2015	18.58	18.2085	12.44	16.2705	0.48	16.1709	0.14
2003	18.9269	20.7473	9.62	19.5247	3.16	20.1876	6.66	20.7374	9.57	19.8922	5.10	18.1776	3.96	18.7352	1.01
2004	22.0738	22.6240	2.49	21.2410	3.77	21.7822	1.32	22.3961	1.46	21.7410	1.51	21.4992	2.60	22.2328	0.72
2005	25.0835	24.4723	2.44	23.1995	7.51	23.5018	6.31	24.1875	3.57	23.7551	5.30	24.4693	2.45	24.9578	0.50
2006	27.5134	26.3039	4.40	25.4343	7.56	25.3562	7.84	26.1221	5.06	25.9344	5.74	26.7641	2.72	27.6156	0.37
2007	29.9271	28.1686	5.88	27.9844	6.49	27.3561	8.59	28.2115	5.73	28.2789	5.51	29.1779	2.50	29.7258	0.67
2008	30.6455	30.0677	1.89	30.8943	0.81	29.5129	3.70	30.4681	0.58	30.7886	0.47	30.6731	0.09	30.8884	0.79
2009	32.1336	32.1466	0.04	34.2147	6.48	31.8390	0.92	32.9051	2.40	33.4635	4.14	30.8443	4.01	32.0259	0.34
MAPE(%)			6.87		5.81		9.87		9.32		6.50		2.85		1.08
2010	34.3601	34.3392	0.06	38.0036	10.60	34.3477	0.04	35.5370	3.43	36.3037	5.66	30.3828	11.58	32.0774	6.64
2011	37.0163	36.5841	1.17	42.3270	14.35	37.0534	0.10	38.3795	3.68	39.3090	6.19	30.0486	18.82	32.6539	11.79
2012	38.1515	38.8779	1.90	47.2603	23.88	39.9715	4.77	41.4493	8.64	42.4796	11.34	30.7685	19.35	32.5588	14.66
2013	39.4794	41.2164	4.40	52.8897	33.97	43.1187	9.22	44.7647	13.39	45.8154	16.05	30.8124	21.95	32.7597	17.02
2014	40.0299	43.5957	8.91	59.3132	48.17	46.5132	16.20	48.3453	20.77	49.3164	23.20	30.6286	23.49	32.7057	18.30
MAPE(%)			**3.29**		26.19		6.06		9.98		12.49		19.04		13.68

[a] The power and background coefficient are 0.01 and 0.5. [b] The optimal background coefficient is 0.36. [c] The coefficient of determination is 0.9749. [d] The model type is ϵ-SVR with the embedding dimension equal to 2 and the kernel type being radial basis. [e] The model type is feed-forward neural network with 2 lagged inputs and 3 nodes in the only hidden layer.

6.3. Comparison with other models

The proposed model is compared with the grey models including the linear discrete one NDGM(1,1) (Xie et al., 2013a), the nonlinear continuous ones NBGM(1,1) (Chen et al., 2008) and GPMB(1,1,N) (Wei et al., 2018), and also some other models including the QPR (quadratic polynomial regression), SVR (support vector regression) and NNAR (neural network autoregression) (Hastie et al., 2013). In order to ensure the comparability of modeling results, the time series in the period from 1990 to 2005 is used for fitting and the rest for evaluating the predicting performance. The calculation results together with the absolute percentage errors (APEs) in two data sets ATEC_{cvc} and ATEC_{cec} are shown in Tables 3 and 4, respectively.

In comparison with the grey models, all their MAPEs are less than 10%, and DGPM and NDGM(1,1) having small differences in MAPE, outperform NBGM(1,1) and GPMB(1,1,0) in terms of the fitting performance. But in the predicting period from 2010 to 2014, their MAPEs differ greatly from each other. It is worth noting here that NDGM(1,1) has a little higher fitting accuracy but much worse predicting performance than DGPM. From the property analysis in subsection 2.4, NDGM(1,1) is equivalent to DGPM(1,1,1) except for the initial value optimization, and the modeling results in subsection 6.2 indicate that the best model structure is $x^{(0)}(k) = \alpha x^{(1)}(k-1) + \beta k^2 + \gamma$ rather than that of DGPM(1,1,1) $x^{(0)}(k) = ax^{(1)}(k-1) + bk + c$. It means that NDGM(1,1) does not recognize the true pattern hidden in the time series and thus results in poor predictions, although optimizing the initial value by least squares again in NDGM(1,1), could reduce the fitting errors.

Table 4: Comparison of the fitted results (from 1990 to 2009) and predicted results (from 2010 to 2014) by using various grey and other models in the data set $ATEC_{cec}$.

Year	Actual	DGPM		NDGM(1,1)		NBGM(1,1) [a]		GPMB(1,1,0) [b]		QPR [c]		SVR [d]		NNAR [e]	
	values	Values	APE	Values	APE	Values	APE	Values	APE	Values	APE	Values	APE	Values	APE
1990	9.8703	—	—	10.2751	4.10	—	—	10.4117	5.48	11.2948	14.43	—	—	—	—
1991	10.3783	11.1164	7.11	10.2792	0.96	8.3533	19.51	8.5854	17.28	10.9973	5.96	—	—	—	—
1992	10.9170	10.8847	0.30	10.6466	2.48	9.0479	17.12	9.2731	15.06	10.8710	0.42	12.1680	11.46	11.1042	1.71
1993	11.5993	10.8547	6.42	11.0658	4.60	9.7864	15.63	10.0158	13.65	10.9159	5.89	12.1223	4.51	11.5108	0.76
1994	12.2737	11.0147	10.26	11.5443	5.94	10.5774	13.82	10.8179	11.86	11.1319	9.30	12.2573	0.13	12.1244	1.22
1995	13.1176	11.3652	13.36	12.0904	7.83	11.4273	12.89	11.6844	10.93	11.5190	12.19	12.6289	3.73	12.7382	2.89
1996	13.5192	11.8925	12.03	12.7136	5.96	12.3417	8.71	12.6202	6.65	12.0773	10.67	13.3248	1.44	13.7006	1.34
1997	13.5909	12.6326	7.05	13.4248	1.22	13.3265	1.95	13.6310	0.29	12.8067	5.77	14.0707	3.53	13.9286	2.48
1998	13.6184	13.6124	0.04	14.2364	4.54	14.3874	5.65	14.7227	8.11	13.7073	0.65	14.4043	5.77	13.8566	1.75
1999	14.0569	14.8355	5.54	15.1628	7.87	15.5309	10.49	15.9019	13.13	14.7790	5.14	14.4769	2.99	13.8694	1.33
2000	14.6964	16.2684	10.70	16.2199	10.37	16.7635	14.07	17.1755	16.87	16.0219	9.02	14.8100	0.77	14.6103	0.59
2001	15.5547	17.8948	15.04	17.4264	12.03	18.0925	16.32	18.5511	19.26	17.4359	12.09	15.6583	0.67	15.5874	0.21
2002	16.9577	19.6968	16.15	18.8033	10.88	19.5255	15.14	20.0369	18.16	19.0211	12.17	16.9301	0.16	16.9976	0.24
2003	19.7083	21.6299	9.75	20.3746	3.38	21.0708	6.91	21.6417	9.81	20.7774	5.42	18.9268	3.97	19.4852	1.13
2004	23.0281	23.5846	2.42	22.1680	3.74	22.7373	1.26	23.3751	1.51	22.7049	1.40	22.2484	3.39	23.2098	0.79
2005	26.1369	25.5143	2.38	24.2146	7.35	24.5346	6.13	25.2472	3.40	24.8035	5.10	25.4004	2.82	26.0254	0.43
2006	28.6467	27.4363	4.23	26.5503	7.32	26.4730	7.59	27.2693	4.81	27.0732	5.49	27.8615	2.74	28.7178	0.25
2007	31.1442	29.3993	5.60	29.2159	6.19	28.5636	8.29	29.4534	5.43	29.5141	5.23	30.3590	2.52	30.9440	0.64
2008	32.0611	31.4045	2.05	32.2580	0.61	30.8185	3.88	31.8124	0.78	32.1262	0.20	31.9388	0.38	32.3625	0.94
2009	33.6126	33.5804	0.10	35.7297	6.30	33.2506	1.08	34.3603	2.22	34.9094	3.86	32.1234	4.43	33.4737	0.41
MAPE(%)			6.87		5.68		9.81		9.23		6.52		3.08		1.06
2010	36.0648	35.8755	0.52	39.6919	10.06	35.8738	0.53	37.1123	2.90	37.8638	4.99	31.5950	12.39	33.7780	6.34
2011	38.7043	38.2318	1.22	44.2137	14.23	38.7032	0.00	40.0847	3.57	40.9893	5.90	31.2048	19.38	34.2314	11.56
2012	40.2138	40.6416	1.06	49.3742	22.78	41.7551	3.83	43.2951	7.66	44.2859	10.13	32.0433	20.32	34.2817	14.75
2013	41.6913	43.1007	3.38	55.2636	32.55	45.0470	8.05	46.7627	12.16	47.7537	14.54	32.1154	22.97	34.3843	17.53
2014	42.5806	45.6049	7.10	61.9848	45.57	48.5977	14.13	50.5081	18.62	51.3927	20.69	31.9055	25.07	34.3982	19.22
MAPE(%)			**2.66**		25.04		5.31		8.98		11.25		20.03		13.88

[a] The power and background coefficient are 0.01 and 0.5. [b] The optimal background coefficient is 0.36. [c] The coefficient of determination is 0.9752. [d] The model type is ϵ-SVR with the embedding dimension equal to 2 and the kernel type being radial basis. [e] The model type is feed-forward neural network with 2 lagged inputs and 3 nodes in the only hidden layer.

In comparison with the other three models, QPR obtains almost the same fitting performance with DGPM, but performs poorly in predicting. Comparing the model structures, QPR does not consider the autocorrelation in time series, resulting in the poor predicting performance. With regard to SVR [1] and NNAR [2], both of them obtain too high fitting accuracy to be true, and the prediction results indicate that they both behave badly in the period from 2010 to 2014. It is likely that the small size of modeling samples leads to the over-fitting (Hastie et al., 2013), although we have tried our best to avoid this by simplifying the model structures in modeling process.

Overall, the DGPM models in the data sets $ATEC_{cvc}$ and $ATEC_{cec}$, have the minimal MAPE values of 3.29% and 2.68% in the period from 2010 to 2014, respectively, indicating the highest predicting accuracy in these two case studies.

6.4. Short discussion

All the above models show an interesting phenomenon that the errors of predicted values tend to increase with the increasing of prediction step. In DGPM models, especially, the APEs in 2014 are as much as 8.91% and 7.10%, indicating that long-term and multi-step prediction should be carefully evaluated.

It can be seen from the optimal model structure in subsection 6.2 and Theorem 3 that the pattern hidden in the annual total energy consumption data roughly reflects the quasi-exponential growth, linear adjustments and uncertainty shocks in reality, and also is in line with the characteristics of the time series in Figure 3. Additionally, Tables 3 and 4 show that all these models, except for SVR and NNAR, have higher fitting errors in the two periods of 1994–1996 and 2000–2002, but only the proposed model behave well,

[1] Support vector regression is implemented by using the function `svm` in **R** package `e1071`. URL: https://mirrors.ustc.edu.cn/CRAN/web/packages/e1071/e1071.pdf

[2] Neural network autoregression is implemented by using the function `nnetar` in **R** package `forecast`. URL: https://mirrors.ustc.edu.cn/CRAN/web/packages/forecast/forecast.pdf

illustrating that the proposed model and algorithm are robust to uncertainty shocks in some sense. In fact, the uncertainty shocks, such as the energy-saving and emission-reduction policy (in The Eleventh Five-year Plan) introduced by the Chinese Government in 2006 and Beijing Olympic Games in 2008, changed the original values in 2008 and 2009 (there exists two- and one-year delays) and thus destroy the consistency of data. In this regard, any predictors are subject to errors, although these events should be accommodated within the fairly large uncertainty bounds (Young, 2018).

7. Conclusions

In this study, a novel discrete grey polynomial model is proposed which presents a unified representation for a family of univariate discrete grey models including the popular homogeneous and non-homogeneous ones. By simulating the original time series directly, the proposed model avoids the two-step parameter estimation and the unnecessary inverse accumulating generation, and makes property analysis simple by introducing matrix decomposition. Furthermore, a data-based selection algorithm is presented to search the optimal model structure adaptively, and then large-scale simulations and two real data sets are employed to test the robustness and performance. The results demonstrate the effectiveness and applicability of the proposed model compared with the alternative models.

There are several interesting directions for extending the present work. First, the data-based structure selection algorithm is a discrete process in this study. It would be interesting to examine how a continuous shrinkage method performs, where the estimates of model parameters can be obtained as

$$\underset{\alpha,\beta_0,\cdots,\beta_N}{\arg\min} \left\{ \sum_{k=2}^{n} \left[x^{(0)}(k) - \alpha x^{(1)}(k-1) - \sum_{j=0}^{N} \beta_j k^j \right]^2 + \lambda_1 |\alpha| + \lambda_2 \sum_{j=0}^{N} |\beta_i| \right\}$$

where $\lambda_1, \lambda_2 > 0$ are hyper-parameters. Next, the forcing term in the proposed model can be actually viewed as a polynomial basis expansion, and inspired by this one important direction would be to introduce the kernel method, that is, $x^{(0)}(k) = \varpi x^{(1)}(k-1) + \boldsymbol{\omega}^{\mathrm{T}} \boldsymbol{b}(k)$, where $\boldsymbol{\omega}$ is the unknown weight vector and $\boldsymbol{b}(k)$ is the unknown basis function vector. This is straightforward but tedious because of the semi-parametric characteristic. Last, efforts to improve the accuracy of the proposed method, such as introducing the rolling and moving mechanism, may be one of the main future studies.

Acknowledgments

The authors appreciate the editor and anonymous referees for their insightful comments and suggestions. This work was supported by National Natural Science Foundation of China (Grants 71671090 and 71871117), Joint Research Project of National Natural Science Foundation of China and Royal Society of UK (Grant 71811530338), Fundamental Research Funds for Central Universities of China (Grant NP2018466), and Qinglan Project for excellent youth or middle-aged academic leaders in Jiangsu Province, China.

References

Altay, N., & Iii, W. G. G. (2007). OR/MS research in disaster operations management. *European Journal of Operational Research*, *175*, 475–493.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control, 3rd ed.*. Prentice Hall, Englewood Cliffs, New Jersey.

Chen, C. I., Chen, H. L., & Chen, S. P. (2008). Forecasting of foreign exchange rates of taiwan's major trading partners by novel nonlinear Grey Bernoulli model NGBM(1,1). *Communications in Nonlinear Science & Numerical Simulation*, *13*, 1194–1204.

Chen, C. I., & Huang, S. J. (2013). The necessary and sufficient condition for GM(1,1) grey prediction model. *Applied Mathematics & Computation*, *219*, 6152–6162.

Evans, M. (2014). An alternative approach to estimating the parameters of a generalised grey verhulst model: An application to steel intensity of use in the UK. *Expert Systems with Applications*, *41*, 1236–1244.

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The Elements of Statistical Learning: Prediction, Inference and Data Mining, 2nd ed.*. Springer-Verlag, New York.

L, L. Y., Medo, M., Chi, H. Y., Zhang, Y. C., Zhang, Z. K., & Zhou, T. (2012). Recommender systems. *Physics Reports*, *519*, 1–49.

Li, S. J., Ma, X. P., & Yang, C. Y. (2018). A novel structure-adaptive intelligent grey forecasting model with full-order time power terms and its application. *Computers & Industrial Engineering*, *120*, 53–67.

Liu, J., Xiao, X., Guo, J., & Mao, S. (2014). Error and its upper bound estimation between the solutions of GM(1,1) grey forecasting models. *Applied Mathematics & Computation*, *246*, 648–660.

Liu, P. L., & Shyr, W. J. (2005). Another sufficient condition for the stability of grey discrete-time systems. *Journal of the Franklin Institute*, *342*, 15–23.

Liu, S. F., Yang, Y. J., & Forrest, J. (2017). *Grey Data Analysis: Methods, Models and Applications*. Springer-Verlag, Singapore.

Long, X., Wei, Y., & Long, Z. (2014). Discrete verhulst model based on a linear time-varying. *Grey Systems: Theory & Application*, *4*, 299–310.

Luo, D., & Wei, B. L. (2017). Grey forecasting model with polynomial term and its optimization. *Journal of Grey System*, *29*, 58–69.

Ma, X., & Liu, Z. B. (2016). Research on the novel recursive discrete multivariate grey prediction model and its applications. *Applied Mathematical Modelling*, *40*, 4876–4890.

Shaikh, F., Ji, Q., Shaikh, P. H., Mirjat, N. H., & Uqaili, M. A. (2017). Forecasting China's natural gas demand based on optimised nonlinear grey models. *Energy*, *140*, 941–951.

Tabaszewski, M., & Cempel, C. (2015). Using a set of GM(1,1) models to predict values of diagnostic symptoms. *Mechanical Systems & Signal Processing*, *52-53*, 416–425.

Tim, H., Marcus, O., & William, R. (1996). Neural network models for time series forecasts. *Management Science*, *42*, 1082–1092.

Wang, Y., Liu, Q., Tang, J., Cao, W., & Li, X. (2014). Optimization approach of background value and initial item for improving prediction precision of GM(1,1) model. *Journal of Systems Engineering & Electronics*, *25*, 77–82.

Wang, Z. X., Hipel, K. W., Wang, Q., & He, S. W. (2011). An optimized NGBM(1,1) model for forecasting the qualified discharge rate of industrial wastewater in China. *Applied Mathematical Modelling*, *35*, 5524–5532.

Wei, B. L., Xie, N. M., & Hu, A. Q. (2018). Optimal solution for novel grey polynomial prediction model. *Applied Mathematical Modelling*, *62*, 717–727.

Wu, L. F., Liu, S. F., Cui, W., Liu, D. L., & Yao, T. X. (2014). Non-homogenous discrete grey model with fractional-order accumulation. *Neural Computing & Applications*, *25*, 1215–1221.

Wu, L. F., Liu, S. F., Yang, Y. J., Ma, L. H., & Liu, H. X. (2016). Multi-variable weakening buffer operator and its application. *Information Sciences*, *339*, 98–107.

Wu, L. F., Liu, S. F., Yao, L. G., & Yan, S. L. (2013a). The effect of sample size on the grey system model. *Applied Mathematical Modelling*, *37*, 6577–6583.

Wu, L. F., Liu, S. F., Yao, L. G., Yan, S. L., & Liu, D. L. (2013b). Grey system model with the fractional order accumulation. *Communications in Nonlinear Science & Numerical Simulation*, *18*, 1775–1785.

Xiao, X. P., Guo, H., & Mao, S. H. (2014). The modeling mechanism, extension and optimization of grey GM(1,1) model. *Applied Mathematical Modelling*, *38*, 1896–1910.

Xie, N. M., & Liu, S. F. (2009). Discrete grey forecasting model and its optimization. *Applied Mathematical Modelling*, *33*, 1173–1186.

Xie, N. M., & Liu, S. F. (2015). Interval grey number sequence prediction by using non-homogenous exponential discrete grey forecasting model. *Journal of Systems Engineering & Electronics*, *26*, 96–102.

Xie, N. M., Liu, S. F., Yang, Y. J., & Yuan, C. Q. (2013a). On novel grey forecasting model based on non-homogeneous index sequence. *Applied Mathematical Modelling*, *37*, 5059–5068.

Xie, N. M., & Wang, R. Z. (2017). A historic review of grey forecasting models. *Journal of Grey System*, *29*, 1–29.

Xie, N. M., Zhu, C. Y., Liu, S. F., & Yang, Y. J. (2013b). On discrete grey system forecasting model corresponding with polynomial time-vary sequence. *Journal of Grey System*, *25*, 1–18.

Xu, J., Tan, T., Tu, M., & Qi, L. (2011). Improvement of grey models by least squares. *Expert Systems with Applications*, *38*, 13961–13966.

Yao, T. X., Liu, S. F., & Xie, N. M. (2009). On the properties of small sample of GM(1,1) model. *Applied Mathematical Modelling*, *33*, 1894–1903.

Young, P. C. (2018). Data-based mechanistic modelling and forecasting globally averaged surface temperature. *International Journal of Forecasting*, *34*, 314–335.

Zeng, B., Liu, S. F., & Xie, N. M. (2010). Prediction model of interval grey number based on DGM(1,1). *Journal of Systems Engineering & Electronics*, *21*, 598–603.

Zeng, B., Meng, W., & Tong, M. (2016). A self-adaptive intelligence grey predictive model with alterable structure and its application. *Engineering Applications of Artificial Intelligence*, *50*, 236–244.

Zeng, L. (2018). A gray model for increasing sequences with nonhomogeneous index trends based on fractionalorder accumulation. *Mathematical Methods in the Applied Sciences*, *41*, 1–14.

Zhang, J., Chen, C. S., & Zeng, B. (2015). Demand forecasting of emergency medicines after the massive earthquake — a grey discrete Verhulst model approach. *Journal of Grey System*, *27*, 234–248.

Zhao, Z., Wang, J., Zhao, J., & Su, Z. (2012). Using a grey model optimized by differential evolution algorithm to forecast the per capita annual net income of rural households in China. *Omega*, *40*, 525–532.