文章编号:1007－6735(2023)03－0271－10　DOI: 10.13255/j.cnki.jusst.20211217001

# 体检指标健康预警的灰色–时序组合模型

朱人杰[1,2]，　叶春明[1]
（1. 上海理工大学 管理学院，上海 200093；2. 同济大学附属东方医院，上海 200120）

**摘要：**对于个体健康体检数据而言，传统的以大样本为基础的数学模型无法满足体检数据的建模需求。基于个体体检数据特征分析，首先构建适用于个体体检指标健康预警的近似非齐次指数序列的改进离散灰色模型。其次，为降低单个模型预测精度的有限性，利用方差倒数法为离散灰色模型和差分自回归移动平均模型赋权重，在模型误差平方和达到最小时取得最佳的权重值。从而将两个模型的预测结果进行组合，实现对健康指标的建模与趋势分析，及时掌握个体健康指标的变化并发现潜在的疾病隐患。预测模型在实验数据集上的相对模拟误差与最优基准模型相比有所下降，表明灰色–时序组合模型具有更高的模拟精度，解决了传统的依据单次体检指标进行静态分析的弊端以及单个模型预测结果的局限性，更加关注个体差异，能有效提升健康预警的效果。

**关键词：**灰色–时序组合模型；体检指标；离散灰色模型；差分自回归移动平均模型；健康预警
**中图分类号：**TP 931　　**文献标志码：**A

# Grey time series combination model for health warning of physical examination indexes

ZHU Renjie[1,2],　YE Chunming[1]
(1. *Business School, University of Shanghai for Science and Technology, Shanghai 200093, China*; 2. *Shanghai Easthospital Affiliated to Shanghai Tongji University, Shanghai 200120, China*)

**Abstract:** For individual health examination data, the traditional mathematical model based on large samples can not meet the modeling requirements of physical examination data. Based on the analysis of the characteristics of individual physical examination data, an improved discrete grey model of approximately non-homogeneous index series suitable for individual physical examination indicator health warning was first constructed. Secondly, in order to reduce the limitation of the prediction accuracy of a single model, the inverse variance method was used to assign weights to the discrete grey model and the differential autoregressive moving average model, and the best weight value was obtained when the sum of squares of the model errors reached the minimum. Thus, the prediction results of the two models were combined to achieve the modeling and trend analysis of health indicators, timely grasp the changes of individual health indicators and discover potential disease hazards. The relative simulation error of the prediction model on the experimental data set decreases in comparison with the

**收稿日期:** 2021−12−17
**基金项目:** 上海市哲学社会科学一般项目 (2022BGL010)；国家自然科学基金资助项目（71840003）
**第一作者:** 朱人杰（1982−），男，博士研究生．研究方向：智慧医疗、运营管理．E-mail：171910085@st.usst.edu.cn
**通信作者:** 叶春明（1964−），男，教授．研究方向：管理科学与工程、工业工程、技术经济与管理．E-mail：yechm6464@163.com

optimal benchmark model, which indicates that the grey time series combination model has higher simulation accuracy. The shortcomings of traditional static analysis based on single physical examination indicators and the limitations of single model prediction results are solved. Individual differences are emphasized, and the effect of health warning can be effectively improved.

**Keywords:** *grey time series combination model*; *physical examination index*; *discrete grey model*; *differential autoregressive integrated moving average model*; *health warning*

随着时代的发展、社会需求和疾病谱的改变，以预防为主的大健康理念逐渐深入人心，民众健康预防的意识也逐渐增强，定期体检演变为一种健康生活习惯和社会趋势。健康体检产生的个体体检数据，可以帮助医生和体检者分析个体自身生理状况和潜在的疾病隐患。但是，医生对体检者身体状况的诊断，或者体检者对自身健康情况的判断大多是基于单次体检结果的高低对比，无法发现个体体检指标在不同时期的动态变化趋势。因此，分析个体体检指标的发展规律和变化趋势，发现体检者的潜在疾病隐患，从而提前采取预防和治疗措施，降低个体未来的患病风险，对于保障个体健康具有重大的现实意义。

灰色模型对于“少数据”、“贫信息”的样本具有较高的预测精度，能够通过研究对象有限的数据，挖掘出数据发展规律和新信息，从而实现对序列未来值的预测[1-2]。在疾病预测方面，灰色预测模型主要用于发病率、死亡率的预测[3-4]。而其他典型预测方法虽然在疾病预测中发挥了重要的作用，但是各类模型的适用范围有所差异。时间序列模型通过将疾病数据随时间推移形成的序列视为一个随机序列，并用一定的数学模型来近似拟合这个序列，常用的时间序列模型为ARIMA（autoregressive integrated moving average model）模型[5-6]。基于概率论的马尔可夫链模型通常是基于系统现在的状态来预测系统未来可能存在的状态，例如刘琼等[2]利用隐马尔科夫模型对乙肝发病数量时间序列进行预测[7]。随着神经网络的发展，BP神经网络模型也被大量应用于疾病预测中，并且在疾病预测中具有较好的识别效果[8-9]。多元回归模型常用于传染病发病率的趋势预测，建模过程中应用直线或曲线拟合原始传染病数据，用数字和等式来表达传染病的流行规律[10-13]。近年来，国内外学者将灰色模型与其他模型进行组合，融合多个模型的优势，开展疾病预测研究。王永斌等[14]将灰色模型和广义回归神经网络模型相结合，预测我国尘肺病发病人数。严薇荣等[15]在进行伤寒副伤寒发病率预测时，将GM(1,1)模型和Markov模型进行组合得到新的预测模型，提高了传染病发病率的预测精度。时冬青等[16]综合GM(1,1)模型和马尔可夫链进行预测，实验结果表明组合模型在职业病预测中的高预测精度。

目前，对于个体体检指标的研究主要集中于两个方面：一方面是分析个体体检指标对于疾病诊断的影响或对疾病的预测价值[17-18]；另一方面是在疾病风险预测中，将多个或群体健康体检指标作为预测特征来预测疾病发病率或患病情况[19-20]。然而，针对个体体检指标未来发展趋势预测的研究还较少。通过上述分析可以发现，以上研究大多采用群体健康指标数据集开展疾病预测，而对于个体健康体检指标的预测较少，并且个体健康体检数据的特征也增加了个体健康指标预测的难度。为此，需要构建有效的个体健康指标预测模型，以期准确预测体检指标未来变化趋势或范围，实现个体健康状况的有效预警管理。基于上述分析，考虑体检指标数据为小样本数据，并且更偏向于是一个非齐次指数序列，为提高模型的泛化性和准确性，本文构建了一个离散灰色模型。同时，为提高预测精度，将ARIMA模型和灰色模型进行组合预测，从而充分利用各个模型的优势。

## 1 个体健康体检指标特征分析

随着人们对于健康和自我保健追求的愈加强烈，健康消费市场迅猛发展，个人定期健康体检已成为常态。个人在医疗机构进行体检，得到各类身体指标检查数据。这些具有时间间隔的数据汇总后形成了时间序列，对这些时间序列数据进行数据分析和预测，可以有效地辅助医生和患者了解当前身体状况和指标的未来变化趋势，帮助

人们提前采取应对策略，做好疾病预防。

由于体检指标时间序列数据有其独有的特征，在构建时间序列预测模型时有必要基于其特征进行设计。以单个体检指标$m$为例，指标$m$在时间跨度$1\sim n$之间的检查结果构成一个时间序列$X_m=(x_m(t_1),x_m(t_2),\cdots,x_m(t_n))$。单个体检指标时间序列具有如下特征：

**a.** 数据量小。

随着时间变化，个人健康状况受年龄变化、外界环境等因素影响，使得体检指标具有阶段性和时效性。通常来说，极早期的体检指标对于分析个人当前身体健康状况的可用价值较低，许多体检数据集中仅保留体检者最近 6～8 年的体检指标数据。因此，体检指标时间序列$X_m=(x_m(t_1),x_m(t_2),\cdots,x_m(t_n))$的样本数量非常有限，一般取样本个数介于 6～10 之间。

**b.** 数据的不确定性。

个人体检指标数值常受到生理状况、心理变化、外界环境等多方面因素的影响，甚至由于测量仪器、检测技术水平的参差不齐也会导致指标数据的不准确。所以个体在进行体检时，总会对异常指标进行多次“复查”，将多次体检结果的可能值或取值范围作为最终检查结果。这导致了体检指标序列的区间出现不确定或离散不确定的情况。

**c.** 时间间隔不一致。

时间序列$X_m=(x_m(t_1),x_m(t_2),\cdots,x_m(t_n))$的时间间隔计算公式一般为$\Delta t=t_{k+1}-t_k$，$k=1,2,\cdots,n-1$，当$\Delta t\neq$常数时，将时间序列$X_m$称为非等时距序列。现实生活中，由于各种因素导致个体未能按期进行健康体检，从而导致体检时间序列数据集中缺失某一时间段的数据，出现时间“断层”问题。

**d.** 数据类型异构。

体检指标数据类型异构是指时间序列$X_m$中不同体检指标具有不同的数据类型。举例来说，时间序列$X_m$中可能存在某一元素数据类型是一个区间值，某一元素数据类型为离散灰数，还有元素数据类型为实数，这就使得$X_m$具有数据类型异构的特征。

**e.** 数据具有上下波动性。

体检指标受到自身以及外部等多个因素制约，从而使得单个个体体检时间序列并非呈现明显的单调递变或恒定不变的规律，通常是在一定数值范围内表现出反复的上下波动的特征。

## 2　灰色–时序组合预测模型 NDGM-ARIMA

### 2.1　改进 GM(1,1) 模型——NDGM(1,1)

由于体检指标数据是一个数据量少的小样本数据集，通常数据量级在几至几十。而灰色模型 GM(1,1) 对于“少数据”、“贫信息”的样本具有较高的预测精度。因此，本文考虑使用灰色模型 GM(1,1)。GM(1,1) 模型是灰色系统理论中经典的预测模型，模型的基本思路是利用原始数据得到一组原始数据序列，对原始数据序列进行累加生成新的数据序列，以此来削弱原始数据的随机性，突出和增强原始数据的规律性，实现对原始数据未来变化规律的模糊预测。

GM(1,1) 具体实现步骤如下：

**步骤 1**　设原始数据构成的序列为$X^{(0)}$，对原始序列进行一次累加生成 (1-AGO) 得到新的数据序列$X^{(1)}$。

**步骤 2**　构建新生成序列$X^{(1)}$的紧邻均值生成序列，记为$Z^{(1)}$。由此得到 GM(1,1) 模型的灰色微分方程$x^{(0)}(k)+az^{(1)}(k)=b$。

**步骤 3**　基于最小二乘原理，可得到参数$a,b$满足的条件为$\hat{\boldsymbol{h}}=(a,b)^{\mathrm{T}}=(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}$，矩阵$\boldsymbol{B}$是构造累加矩阵，向量$\boldsymbol{Y}$为常数项向量。

**步骤 4**　由序列$X^{(0)},X^{(1)},Z^{(1)}$可得到 GM(1,1) 模型的白化微分方程，将 GM(1,1) 模型白化方程的解称为时间响应函数。

**步骤 5**　求解得到白化微分方程的时间响应序列后，通过累减生成还原得到原始序列为$\hat{x}^{(0)}(k+1)=\hat{x}^{(1)}(k+1)-\hat{x}^{(1)}(k)$，即灰色 GM(1,1) 的预测方程表达式，对其进行求导还原就可得到序列还原值。

传统的 GM(1,1) 模型是用一阶微分方程对单个变量实现预测的模型，其建模过程主要是利用齐次指数序列来拟合原始数据。因此，GM(1,1) 模型对于具有近齐次指数的原始序列具有较好的拟合与预测性能。但是，现实生活中存在许多不确定因素，绝大部分的时间序列都不符合指数增长规律。对于体检指标序列，这类序列由于数值结果不确定性大、时间间隔不统一导致的数值缺失，以及数据上下波动等原因，使得体检指标序列更符合近似非齐次指数序列变化特征。同时，传统的 GM(1,1) 模型中参数估计方程是离散的，模型

预测方程是连续的，为了解决离散参数估计和连续预测表示之间跳跃所产生的模拟误差，本文借鉴了谢乃明等[21]提出的离散灰色模型DGM(1,1)基本思想，使改进灰色模型的参数估计和模型预测都是离散形式。

结合上述体检序列特征分析和预测模型性能分析，为了构建适用于体检指标序列的预测模型，本文构建一个近似非齐次指数序列的离散GM(1,1)模型(non-homogenous discrete grey model)，简写为NDGM(1,1)模型。

同样地，设原始非负序列为$X^{(0)}:X^{(0)}=(x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(n))$。其中，$x^{(0)}(i)\geqslant 0,i=1,2,\cdots,n$。经过一次累加生成得到新序列$X^{(1)}$：$X^{(1)}=(x^{(1)}(1),x^{(1)}(2),\cdots,x^{(1)}(n))$，从而得到离散灰色模型NDGM(1,1)的表达式为$x^{(1)}(t+1)+ax^{(1)}(t)=bt+c$，则模型的白化微分方程表达式为

$$\frac{\mathrm{d}x^{(1)}}{\mathrm{d}t}+ax^{(1)}=bt+c \tag{1}$$

式中，参数列$\hat{\boldsymbol{h}}=(a,b,c)^{\mathrm{T}}$为NDGM(1,1)模型待求解参数。

求解NDGM(1,1)模型白化方程的时间响应序列，首先公式对应的齐次方程为

$$\frac{\mathrm{d}x^{(1)}(t)}{\mathrm{d}t}+ax^{(1)}(t)=0\Rightarrow\frac{\mathrm{d}x^{(1)}(t)}{\mathrm{d}t}=-ax^{(1)}(t) \tag{2}$$

解出齐次方程的通解为$x^{(1)}(t)=C_1\mathrm{e}^{-at}$。利用常数变易法，令$C_1=f(t)$，则$x^{(1)}(t)=f(t)\mathrm{e}^{-at}$。对$x^{(1)}(t)=f(t)\mathrm{e}^{-at}$两端同时求导后代入式(2)可得

$$f'(t)\mathrm{e}^{-at}-af(t)\mathrm{e}^{-at}=bt+c-ax^{(1)} \tag{3}$$

$$f'(t)=(bt+c)\mathrm{e}^{at} \tag{4}$$

$$f(t)=\int(bt+c)\mathrm{e}^{at}\mathrm{d}t=\frac{b}{a}t\mathrm{e}^{at}-\frac{b}{a^2}\mathrm{e}^{at}+\frac{c}{a}\mathrm{e}^{at}+C \tag{5}$$

将式(5)代入$x^{(1)}(t)=C_1\mathrm{e}^{-at}$中，可知

$$x^{(1)}(t)=\frac{b}{a}t-\frac{b}{a^2}+\frac{c}{a}+C\mathrm{e}^{-at} \tag{6}$$

当$t=1$时，可得$x^{(1)}(1)=\frac{b}{a}t-\frac{b}{a^2}+\frac{c}{a}+C\mathrm{e}^{-a}$，解出$C$的表达式为

$$C=\frac{x^{(1)}(1)-\frac{b}{a}t+\frac{b}{a^2}-\frac{c}{a}}{\mathrm{e}^{-a}} \tag{7}$$

将式(7)代入式(6)得到NDGM(1,1)模型的时间响应序列表达式为

$$x^{(1)}(t+1)=\mathrm{e}^{-a}x^{(1)}(t)+\frac{b}{a}(1-\mathrm{e}^{-a})t+(1-\mathrm{e}^{-a})\left(\frac{c}{a}-\frac{b}{a^2}\right)+\frac{b}{a} \tag{8}$$

则式(8)经过累减还原得到还原式为

$$\hat{x}^{(0)}(t)=\hat{x}^{(1)}(t)-\hat{x}^{(1)}(t-1)=(1-\mathrm{e}^{a})\left(x^{(0)}(1)-\frac{b}{a}+\frac{b}{a^2}-\frac{c}{a}\right)\mathrm{e}^{-a(t-1)}+\frac{b}{a},\quad t=2,3,\cdots,n,\cdots \tag{9}$$

当$t=2,3,4,\cdots,n$时，$\hat{x}^{(0)}(t)$为模型所得拟合值；当$t=n+1,n+2,\cdots$时，$\hat{x}^{(0)}(t)$为模型所得预测值。

令$\alpha=\mathrm{e}^{-a}$，$\beta=\frac{b}{a}(1-\mathrm{e}^{-a})$，$\gamma=(1-\mathrm{e}^{-a})\left(\frac{c}{a}-\frac{b}{a^2}\right)+\frac{b}{a}$，则式(8)可表示为

$$x^{(1)}(t+1)=\alpha x^{(1)}(t)+\beta t+\gamma \tag{10}$$

式(10)的参数列$\hat{\boldsymbol{C}}=(\alpha,\beta,\gamma)^{\mathrm{T}}$，由最小二乘法得到参数的估计值，当式(11)所示的误差平方和达到最小时可求解出参数$\alpha,\beta,\gamma$。

$$S=\sum_{t=1}^{n-1}[x^{(1)}(t+1)-\hat{\alpha}x^{(1)}(t)-\hat{\beta}t-\hat{\gamma}]^2 \tag{11}$$

参数列$\hat{\boldsymbol{C}}=(\alpha,\beta,\gamma)^{\mathrm{T}}$应满足条件$(\alpha,\beta,\gamma)^{\mathrm{T}}=(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}$，其中

$$\boldsymbol{B}=\begin{bmatrix}x^{(1)}(1)&2&1\\x^{(1)}(2)&3&1\\\vdots&\vdots&\vdots\\x^{(1)}(n-1)&n&1\end{bmatrix}\quad\boldsymbol{Y}=\begin{bmatrix}x^{(1)}(2)\\x^{(1)}(3)\\\vdots\\x^{(1)}(n)\end{bmatrix}$$

$a,b,c$的估计值分别为

$$\hat{a}=-\ln\hat{\alpha},\ \hat{b}=\frac{\hat{a}\hat{\beta}}{1-\hat{\alpha}},\ \hat{c}=\frac{\hat{a}\hat{\gamma}-\hat{b}}{1-\hat{\alpha}}+\frac{\hat{b}}{\hat{a}} \tag{12}$$

将参数估计值$\hat{a},\hat{b},\hat{c}$代入式(9)所得的还原式，即可求出原始数据序列的模拟值和预测值。

NDGM(1,1)模型建立后，为了评价模型运行的可行性，需要对模型进行精度检验，本文利用后残差检验法进行检验。记原始序列$X^{(0)}:X^{(0)}=(x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(n))$和残差序列$\varepsilon^{(0)}=(\varepsilon(1),\varepsilon(2),\cdots,\varepsilon(n))=(x^{(0)}(1)-\hat{x}^{(0)}(1),x^{(0)}(2)-\hat{x}^{(0)}(2),\cdots,x^{(0)}(n)-\hat{x}^{(0)}(n))$的方差分别为$S_1^2$，$S_2^2$，计算公式分别为

$$S_1^2=\frac{1}{n-1}\sum_{k=1}^{n}(x^{(0)}(k)-\bar{x}^{(0)})^2 \tag{13}$$

$$S_2^2=\frac{1}{n-1}\sum_{k=2}^{n}(\varepsilon^{(0)}(k)-\bar{\varepsilon}^{(0)})^2 \tag{14}$$

式中：$\bar{x}^{(0)}$表示原始序列的均值，计算公式为$\bar{x}^{(0)}=\frac{1}{n}\sum_{k=1}^{n}x^{(0)}(k)$；$\bar{\varepsilon}^{(0)}$为序列残差均值，且$\bar{\varepsilon}^{(0)}=\frac{1}{n}\sum_{k=2}^{n}\varepsilon^{(0)}(k)$。

后验残差检验法是利用后验差比值$c$和小概率误差$p$进行检验，二者计算方法为

$$c=\frac{S_2}{S_1} \tag{15}$$

$$p=P\left\{0.674\,5S_1>\left|e^{(0)}(k)-\bar{e}^{(0)}\right|\right\} \tag{16}$$

若 NDGM(1,1) 模型满足表 1 所示的模型精度标准，则说明构建的 NDGM(1,1) 模型合格。

**表 1　灰色预测模型精度表**

**Tab.1　Precision of grey prediction model**

| 精度等级 | $p$值 | $c$值 |
|---|---|---|
| 很好 | $p\geqslant 0.95$ | $c\leqslant 0.35$ |
| 合格 | $0.80\leqslant p<0.95$ | $0.35<c\leqslant 0.50$ |
| 勉强合格 | $0.70\leqslant p<0.80$ | $0.50<c\leqslant 0.65$ |
| 不合格 | $p<0.70$ | $c>0.65$ |

## 2.2　ARIMA 模型的构建

将时间序列定义为一组按时间先后顺序排列的数据集合，时间序列预测就是指利用模型分析和处理时间序列，根据时间序列呈现出的规律，构建有效的模型对数据未来发展趋势进行预测。常用于预测平稳时间序列的时间序列模型包括自回归模型 AR($n$)、自回归移动平均模型 ARMA($p$,$q$)、差分自回归移动平均模型 ARIMA($p$,$d$,$q$)。

ARIMA($p$,$d$,$q$) 模型的建模过程为，首先将非平稳时间序列经处理后转化为平稳时间序列，然后将因变量只对其滞后值（阶数）以及随机误差项的现值和滞后值进行回归分析。ARIMA($p$,$d$,$q$) 模型对于短期时间序列预测具有较高的预测精度。其中：AR 表示自回归；MA 表示移动平均；$p$，$q$分别表示 AR，MA 的阶数；$d$表示差分的阶数，一般取值为 1 阶或 2 阶序列达到稳定。ARIMA($p$,$d$,$q$) 模型为

$$\begin{cases}\Phi(B)\nabla^d x_t=\Theta(B)\varepsilon_t\\ \mathrm{E}(\varepsilon_t)=0,\mathrm{var}(\varepsilon_t)=\sigma_\varepsilon^2,\ \mathrm{E}(\varepsilon_t\varepsilon_s)=0,s\neq t\\ \mathrm{E}(x_s\varepsilon_t)=0,\forall s<t\end{cases} \tag{17}$$

式中：$\nabla^d=(1-B)^d$为差分运算；$\{\varepsilon_t\}$表示零均值白噪声序列；$\Phi(B)=1-\phi_1B-\cdots-\phi_pB^p$，$\Theta(B)=1-\theta_1B-\cdots-\theta_pB^p$分别表示模型 ARIMA($p$,$d$,$q$) 的自回归系数多项式和移动平均系数多项式；$B$ 表示延迟算子，并且满足$B^n x_t=x_{t-n}$。

ARIMA($p$,$d$,$q$) 模型的建模包括时间序列预处理、模型识别和定阶、模型检验、模型验证及优化和模型预测 5 个步骤。本文利用 Eviews 软件进行 ARIMA 模型确定和指标预测，具体过程如下：

**a.** 时间序列预处理。适用于 ARIMA($p$,$d$,$q$) 模型的时间序列必须为平稳非白噪声时间序列，对于非平稳时间序列，需进行数据预处理使原始序列满足平稳化和零均值的条件。将实验序列数据录入 Eviews 软件后，通过绘制原始序列的时序图来判断序列的平稳性。若序列是非平稳状态，采用取对数或差分处理等操作进行处理，处理完后进行 ADF 单位根检验序列平稳性。

**b.** 模型识别和定阶。对于模型的识别和定阶本质上就是确定参数$p$，$q$的值，基于数据预处理后的平稳时间序列，计算出实验数据集的自相关系数 ACF 和偏自相关系数 PACF。对预处理后的序列通过 Eviews 软件的 Correlogram 得到序列自相关图和数值，采用 AIC 准则为预测模型的阶数$p$和$q$取合适的值。

**c.** 模型检验。对识别和定阶后的 ARIMA 模型进行参数估计，模型的检验包括参数估计的显著性检验和残差序列的随机性检验，即验证残差之间的独立性。确定 ARIMA 模型各项阶数后，在 Eviews 中创建估计方程式得到 Prob.值，Prob.值若小于 5% 则模型是显著的，可靠性较高。

**d.** 模型的验证和优化。根据模型检验结果对模型的阶数进行调整和优化，使构建出的模型满足显著性检验要求。即若步骤 c 中得到的模型估计结果未通过检验，则返回修改模型阶数$p$和$q$，重新进行检验。

**e.** 模型拟合和预测。利用构建好的 ARIMA 模型对实验时间序列进行拟合，并预测数据未来的趋势。对于检验通过的 ARIMA 模型利用 Eviews 中的 Forecast 模块，在 sample 栏中选择需预测的实验数据进行逐步向前预测。

## 2.3　组合预测模型 NDGM-ARIMA

各类预测模型的研究重点和关注方向都有所不同，因此，对同一个实验数据集进行预测，不同的模型会产生不同的结果。为了提高预测模型的预测精度以及模型的适用性，本文将 NDGM(1,1) 模型和 ARIMA($p$,$d$,$q$) 模型进行组合，简称 NDGM-ARIMA 模型。组合预测模型综合考虑两个模型的预测结果，通过为单个模型的预测结果赋予最佳的权重系数，最大限度地利用多个模型的样本信

息。构建组合模型，也在一定程度上减少了单个预测模型受外界因素的干扰，考虑问题更加全面系统，从而提高模型预测的精度。

本文构建的NDGM-ARIMA组合预测模型用于实现个人体检指标序列的预测，模型具体的表达式为：$\hat{X}(t)=w\hat{G}(t)+(1-w)\hat{A}(t)$。其中：$\hat{G}(t)$表示NDGM(1,1)模型$t$时刻的预测值；$\hat{A}(t)$表示ARIMA模型$t$时刻的预测值；$w$为组合模型权重值，取值范围为$w\in[0,1]$，表示单个模型预测结果的重要程度。

在组合预测模型中，如何恰当地求解出权重系数是关键。确定权重系数常用方法包括算术平均法、最优加权法、方差倒数法等。算术平均法是在对模型重要性缺乏了解时常用的权重选定方法，但是该方法缺乏对单个模型重要性的掌握，对每个模型赋予相同的权重，不分优先顺序使得预测效果不佳。最优加权法需要求解线性或非线性规划，计算复杂并且计算结果有可能为负，在实际应用中具有较大的局限性。方差倒数法则是通过预测模型的误差平方和的计算来反映预测精度，相较于算数平均法和最优加权法，直接应用预测误差平方和更能反映各个模型在组合预测中的重要程度，赋予的权重数值更为合理有效。而且方差倒数法易操作，获得的预测效果好。因此，为求解预测模型最佳的组合权重大小，本文采用方差倒数这一方法。方差倒数的目的是使组合预测模型的误差平方和尽可能小。因此，需要对组合模型中误差平方和大的模型赋较小的权重值，对误差平方和小的模型赋较大的权重值。

采用方差倒数进行组合权重赋值，首先计算出单个预测模型的预测误差平方和。用$e_i$表示第$i$个模型的误差平方和，其计算方式如式(18)所示。

$$e_i=\sum_{t=1}^{n}(x_i-\hat{x}_{ti})^2 \tag{18}$$

式中：$x_i$为原始数据；$\hat{x}_{ti}$为其对应的预测值；$(x_i-\hat{x}_{ti})$为预测误差。

计算出单个模型的误差平方和在全部模型中的占比，这一占比即该模型的权重值大小。利用模型的预测误差得到权重系数的计算公式为

$$w_i=\frac{e_i^{-1}}{\sum_{i=1}^{m}e_i^{-1}} \tag{19}$$

式中，$\sum_{i=1}^{m}w_i^{-1}=1,j=1,2,\cdots,m$。

由式(19)可以发现，当单个模型的误差平方和越大时，获得的权重越小，则模型预测精度越低，预测结果的价值度越低。

## 3 个体体检指标预测实验及结果分析

### 3.1 实验数据集描述

心血管疾病已成为当前社会的一种高发疾病，该类疾病的高危致病因素众多，包括高血压、糖尿病、肥胖、血脂异常、吸烟和过度饮酒等。由相关统计数据可发现，近年来，心血管疾病患者死亡率极高，所以人们必须对此类疾病引起重视，加强自身健康管理。患者通过定期健康体检，可以帮助医生和患者及时了解当前身体状况，发现关键病因信号，提前进行预防和治疗，降低患病的风险。因此，构建适当的预测模型，实现对人体主要健康指标序列的有效预测，具有重要的现实意义。

本文采用天池公开数据集中的心脏病体检数据集进行分析，数据集中包含多名体检者连续多年的体检数值，例如血脂水平中甘油三酯、总胆固醇、高密度脂蛋白胆固醇、低密度脂蛋白胆固醇4项指标和空腹血糖指标等数值。实验选择空腹血糖指标作为实验数据序列，血糖指标是检测心血管疾病和糖尿病的关键指标，同时也是人体健康管理中重要的体检指标，关注血糖值的变化可以有效监测到心血管类疾病。空腹血糖指标的正常取值为3.9～6.1 mmol/L。在4个不同年龄段（20～30岁，30～40岁，40～50岁，50～60岁）中随机选择一名体检者，对4名体检者的空腹血糖指标进行拟合和预测。4名体检者2005—2014年指标的空腹血糖体检时间序列为表2，将4名实验对象样本分别用$X_1$，$X_2$，$X_3$，$X_4$表示。

### 3.2 模型预测结果及分析

为了更加直观地分析组合预测模型的性能，利用ARIMA($p,d,q$)，GM(1,1)，NDGM(1,1)，NDGM-ARIMA组合预测模型4个模型对血糖体检时间序列进行拟合和预测，通过分析各模型的预测值和相对模拟误差$\Delta(t)$来分析组合预测模型的预测性能。相对模拟误差计算公式如下：

$$\Delta(t)=\frac{\left|\hat{x}^{(0)}(t)-x^{(0)}(t)\right|}{x^{(0)}(t)} \tag{20}$$

**表 2　4 名体检者 2005—2014 年空腹血糖体检数据**

**Tab.2　Fasting blood glucose physical examination data of 4 examines from 2005 to 2014**

| 体检年份 | 空腹血糖指标 /(mmol·L$^{-1}$) | | | |
|---|---|---|---|---|
| | $X_1$(27岁) | $X_2$(35岁) | $X_3$(45岁) | $X_4$(56岁) |
| 2005 | 4.26 | 4.04 | 5.09 | 5.47 |
| 2006 | 4.53 | 4.36 | 5.03 | 5.34 |
| 2007 | 4.59 | 4.47 | 4.94 | 5.43 |
| 2008 | 4.48 | 4.64 | 5.73 | 5.73 |
| 2009 | 4.52 | 4.72 | 5.03 | 5.37 |
| 2010 | 4.63 | 4.69 | 5.36 | 5.14 |
| 2011 | 4.41 | 4.89 | 4.78 | 5.59 |
| 2012 | 4.72 | 4.78 | 4.92 | 5.51 |
| 2013 | 4.62 | 4.99 | 5.62 | 5.92 |
| 2014 | 4.63 | 5.16 | 5.78 | 5.49 |

### 3.2.1　ARIMA(2,2,1) 模型预测

实验分别对 4 名体检者空腹血糖时间序列建立相应的 ARIMA 预测模型。例如 45 岁体检者的时间序列由原始序列 $X^{(0)}=(x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(10))=(5.69,5.03,\cdots,5.78)$可知，原始序列是一个非平稳时间序列，首先进行差分处理转化为平稳序列。将数据输入 Eviews 软件中，对原始序列进行 ADF 检验，ADF 检验结果如图 1 所示。可发现当二阶差分时，所有 $t$ 值的绝对值均小于 ADF 检验统计量的绝对值，且 $p$ 值为 0.009 0，小于 0.05，说明原序列已转化为平稳时间序列，则 ARIMA 模型的差分阶数为$d=2$。

| | | t-statistic | Prob.* |
|---|---|---|---|
| Augmented dickey-fuller test statistic | | −4.901 545 | 0.009 0 |
| Test critical values: | 1% level | −4.803 492 | |
| | 5% level | −3.403 313 | |
| | 10% level | −2.841 819 | |

**图 1　ADF 检验**

**Fig.1　ADF inspection**

接着，对模型进行识别，确定模型的 ACF 和 PACF。利用 Eviews 软件 Correlogram 相关图查看序列二阶差分的 ACF 和 PACF 值，得到如图 2 所示的自相关图。

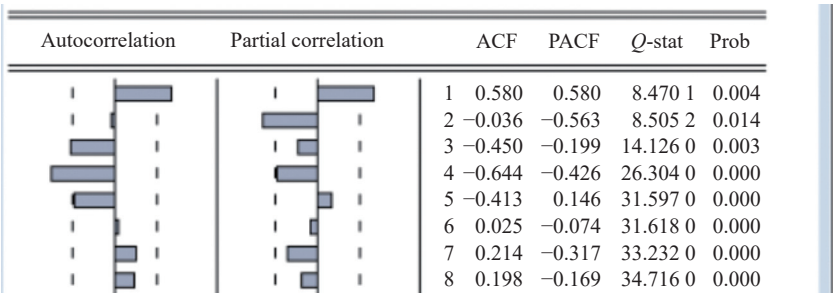


**图 2　序列的自相关系数和偏自相关系数**

**Fig.2　Autocorrelation coefficient and partial autocorrelation coefficient of sequence**

由图 2 可知，时间序列的自相关系数 ACF 在 1 阶截尾，偏自相关系数 PACF 在 2 阶截尾。因此，构建 ARIMA(2,2,1) 模型对空腹血糖体检序列进行预测。之后，在 Eviews 软件中进行建模，采用列表法对 ARIMA 方程进行定义：data c ar(1) ar(2) ma(1)，根据定义后的模型得到 ARIMA(2,2,1) 模型具体的表达式为

$$\hat{X}(t)=0.100\,054+1.577\,592X_{t-1}-0.886\,659X_{t-2}+0.799\,962\,1a_{t-1}+\varepsilon_t \quad (21)$$

对于 27，35，57 岁体检者血糖序列，同样利用 Eviews 软件建立最优的 ARIMA 模型，得到 27 岁体检者血糖序列的时间序列预测模型为 ARIMA(2,1,1)，35 岁对应模型为 ARIMA(3,1,2)，57 岁对应模型为 ARIMA(3,2,2)。

### 3.2.2　NDGM(1,1) 模型预测

同样地，对于 4 个时间序列建立对应的 GM(1,1) 模型和离散灰色模型 NDGM(1,1)。以表 2 所示的 45 岁体检者的血糖数据为具体例子进行建模，可知该体检者空腹血糖指标原始序列为 $X^{(0)}=(x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(10))=(5.09,5.03,\cdots,5.78)$，利用 python 代码建立序列$X^{(0)}$的 GM(1,1) 模型，得到模型参数$a=-0.010\,26$，$b=4.946\,395$，则关于空腹血糖指标预测的 GM(1,1) 模型的时间响应表达式为

$$\begin{cases}\hat{x}^{(1)}(1)=5.69\\ \hat{x}^{(1)}(k+1)=487.794\,8\mathrm{e}^{0.010\,26k}-482.104\,8\end{cases} \quad (22)$$

进一步对优化后的 NDGM(1,1)模型的参数$\hat{\alpha}$，$\hat{\beta}$，$\hat{\gamma}$及$a$，$b$，$c$进行参数估计，计算出具体的数值结

果，$\hat{\alpha}=0.128\,6$，$\hat{\beta}=4.556\,4$，$\hat{\gamma}=5.387\,5$，$a=2.050\,7$，$b=10.723\,4$，$c=5.601\,8$，得到NDGM(1,1)模型为

$$\hat{x}^{(0)}(t)=0.279\,1(1-\mathrm{e}^{2.050\,7})\mathrm{e}^{-2.050\,7(t-1)}+5.229\,1 \quad (23)$$

同理可得：27岁体检者的GM(1,1)模型和NDGM(1,1)模型表达式分别如式(24)和(25)所示；35岁体检者的GM(1,1)模型和NDGM(1,1)模型表达式分别如式(26)和(27)所示；57岁体检者的GM(1,1)模型和NDGM(1,1)模型表达式分别如式(28)和(29)所示。

$$\hat{x}^{(1)}(k+1)=1\,206.11\mathrm{e}^{0.003\,74k}-1\,201.85 \quad (24)$$

$$\hat{x}^{(0)}(t)=0.482\,6(1-\mathrm{e}^{0.249\,6})\mathrm{e}^{-0.249\,6(t-1)}+4.657\,89 \quad (25)$$

$$\hat{x}^{(1)}(k+1)=238.468\,8\mathrm{e}^{0.018\,3k}-234.429 \quad (26)$$

$$\hat{x}^{(0)}(t)=0.420\,6(1-\mathrm{e}^{-0.108})\mathrm{e}^{0.108(t-1)}+3.938\,9 \quad (27)$$

$$\hat{x}^{(1)}(k+1)=953.810\,9\mathrm{e}^{0.005\,62k}-948.341 \quad (28)$$

$$\hat{x}^{(0)}(t)=0.537\,0(1-\mathrm{e}^{-0.945\,5})\mathrm{e}^{0.945\,5(t-1)}+5.457\,1 \quad (29)$$

### 3.2.3 灰色-时间序列组合模型NDGM-ARIMA预测

将4个预测模型ARIMA(2,2,1)，GM(1,1)，NDGM(1,1)和NDGM-ARIMA组合模型分别对4名体检者2005—2014年空腹血糖体检序列进行预测，各个模型对45岁体检者血糖的预测结果如表3所示，4名体检者的整体预测结果如图3所示。利用式(19)的权重系数计算方法确定组合模型的权重系数，得到在对45岁体检者进行预测时，NDGM(1,1)模型和ARIMA(2,2,1)模型的权重系数分别为0.628 6，0.371 4。

**表3 空腹血糖指标的模型拟合结果**

**Tab.3 Model fitting results of fasting blood glucose index**

| 年份 | 空腹血糖 | ARIMA(2, 2, 1)模型 | | GM(1, 1)模型 | | NDGM(1, 1)模型 | | NDGM-ARIMA模型 | |
|---|---|---|---|---|---|---|---|---|---|
| | | 预测值 | 相对误差 | 预测值 | 相对误差 | 预测值 | 相对误差 | 预测值 | 相对误差 |
| 2006 | 5.03 | 4.83 | 0.039 8 | 5.03 | 0.000 1 | 4.99 | 0.008 8 | 5.00 | 0.006 0 |
| 2007 | 4.94 | 5.01 | 0.014 2 | 5.08 | 0.028 8 | 5.10 | 0.032 4 | 5.06 | 0.024 3 |
| 2008 | 5.73 | 5.22 | 0.089 0 | 5.13 | 0.103 9 | 5.18 | 0.096 0 | 5.21 | 0.090 8 |
| 2009 | 5.03 | 5.28 | 0.049 7 | 5.19 | 0.031 4 | 5.20 | 0.033 8 | 5.38 | 0.069 6 |
| 2010 | 5.36 | 5.34 | 0.003 7 | 5.24 | 0.022 1 | 5.26 | 0.018 7 | 5.27 | 0.016 8 |
| 2011 | 4.78 | 5.40 | 0.129 7 | 5.30 | 0.107 8 | 5.31 | 0.110 9 | 5.35 | 0.119 2 |
| 2012 | 4.92 | 5.52 | 0.122 0 | 5.35 | 0.087 4 | 5.52 | 0.122 0 | 5.40 | 0.097 6 |
| 2013 | 5.62 | 5.73 | 0.019 6 | 5.41 | 0.038 2 | 5.73 | 0.019 6 | 5.64 | 0.003 6 |
| 2014 | 5.78 | 5.97 | 0.032 9 | 5.46 | 0.055 2 | 5.88 | 0.017 3 | 5.80 | 0.005 2 |

由图3所示的4名体检者的预测结果曲线和实际数据曲线对比分析可知，论文对于35岁体检者的空腹血糖指标预测结果并非是对比模型中最佳的。这有可能是因为在数据集中，该体检者初始体检年份血糖指标与最终体检年份的指标数值相差较大。由于存在各种外界因素导致的两个体检数据的不准确和差距较大，使得模型的误差较大，从而导致预测精度下降。但是，通过进一步分析35岁体检者空腹血糖指标预测值可以发现，构建的组合模型与最优预测模型二者间的预测值相差极小。同时，组合预测模型在其余3个年龄段的体检者的血糖值拟合上都更接近真实数值，说明了组合模型对于绝大多数的体检数据预测是有效的，也证明了组合模型预测结果的真实性、高可信度。

进一步对45岁体检者血糖指标预测具体数值进行分析。与传统的GM(1,1)模型对比，改进的灰色模型NDGM(1,1)在实验序列上的拟合值虽然存在部分预测值差于GM(1,1)模型，但是从两个模型的平均相对误差来看，NDGM(1,1)模型的平均相对误差为0.050 1，GM(1,1)模型的平均相对误差为0.052 8，NDGM(1,1)模型的平均相对误差小于GM(1,1)。这一实验结果显示，构建的NDGM(1,1)模型在体检指标序列预测上整体的预测效果优于GM(1,1)模型，说明构建的改进灰色预测模型在预测精度上得到了一定程度的提升。

通过NDGM-ARIMA模型与3个单个预测模型的对比，组合模型的拟合值和相对误差都优于单个灰色预测和时间序列模型，这说明组合模型确实适用于健康体检指标的预测，模型的拟合值

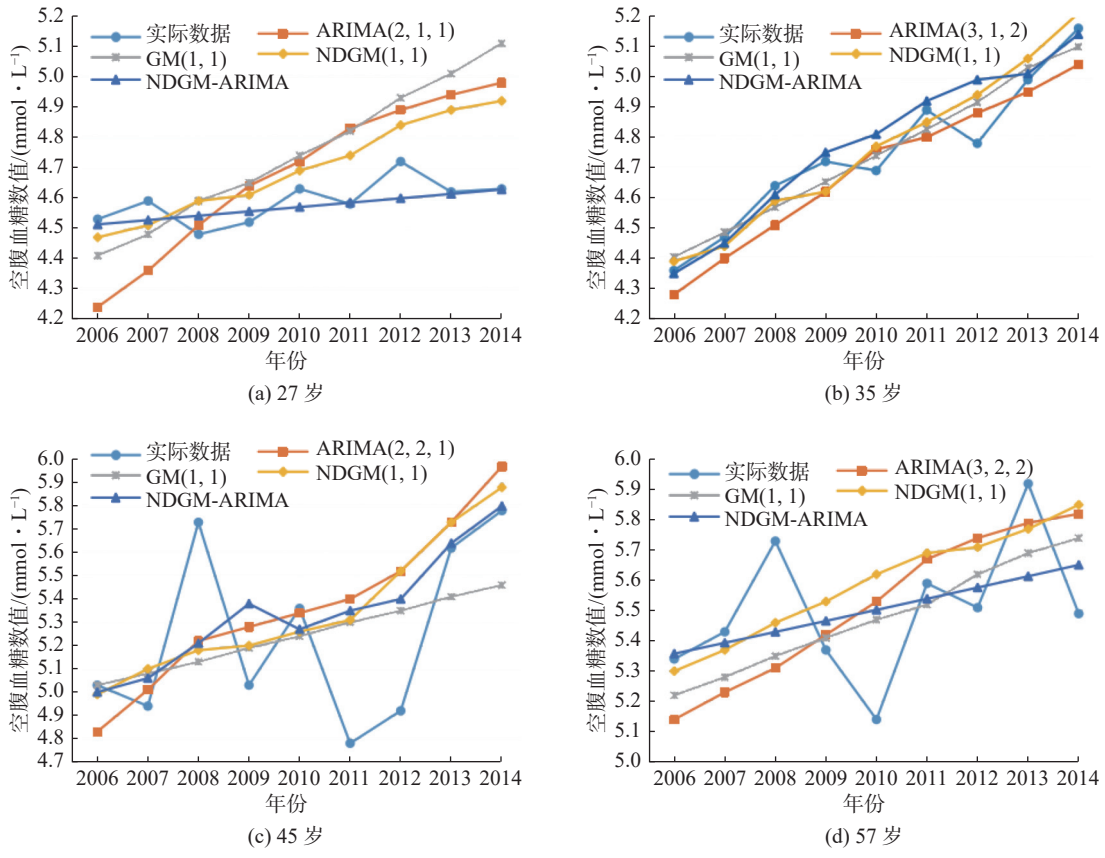


**图 3　4 名体检者血糖指标预测结果**

**Fig.3　Prediction results of blood glucose indexes of 4 physical examiners**

更加接近实际体检数据。另外，这也证明了组合模型能够更好地结合单个 NDGM(1，1) 模型和时间序列模型的优点，在一定程度上克服了单一预测模型的局限性，提高了模型的预测精度。

利用 NDGM-ARIMA 组合模型预测 45 岁体检者 2015—2018 年的血糖数值，预测结果如表 4 所示。

**表 4　NDGM-ARIMA 模型对 2015—2018 年空腹血糖指标预测值**

**Tab.4　NDGM-ARIMA model predicted values of fasting blood glucose indexes from 2015 to 2018**

| 年份 | 血糖预测数值/(mmol·L$^{-1}$) |
|---|---|
| 2015 | 5.87 |
| 2016 | 5.94 |
| 2017 | 6.12 |
| 2018 | 6.31 |

已知空腹血糖的正常范围为 3.6～6.1 mmol/L，结合 2015—2018 年的预测值来分析该体检者身体状况变化趋势。由预测数值可发现，该体检者的空腹血糖指标数值呈现缓慢上升的趋势，预计到 2017 年血糖指标数值将达到 6.12 mmol/L，已经突破人体空腹血糖正常值最大临界值，体检者极有可能患糖尿病等疾病，危害身体健康。因此，由预测结果可以得出，体检者未来几年患糖尿病和心血管疾病的潜在风险很大，必须注意自身糖分的摄入，加强身体日常管理，提前做好预防措施或采取及时的治疗手段。

## 4　结　论

传统体检指标分析仅局限于单次指标数值高低的静态分析，忽略了因个体差异导致的体检数据的动态变化趋势。因此，构建合理有效的数据模型来挖掘体检指标的发展规律，准确预测体检数值的变化趋势和未来取值范围，并基于预测结果对个体健康状况实施预警管理，通过监测人体主要健康指标的变化，及时发现潜在的患病因子或风险因素，进一步采取有效的预防和治疗措施，对于实现个体健康管理具有重要的现实意义。

为了构建适用于个体主要健康体检指标的预测模型，加强模型在体检指标上的预测性能，本文提出一个改进灰色模型和时间序列模型相结合的组合预测模型。首先分析体检指标序列的特

征，考虑到体检指标序列是一个近似非齐次指数序列，以及GM(1,1)模型中的离散和连续之间的误差，构建了一个近似非齐次指数序列的离散灰色模型NDGM(1,1)。其次，为了将单个预测模型的优势结合在一起，论文将时间序列预测模型ARIMA($p$,$d$,$q$)和NDGM(1,1)模型进行组合得到NDGM-ARIMA模型。在尽可能保证组合模型误差平方和达到最小的情况下，为两个模型的预测结果赋予最佳权重系数，并将加权后的结果作为最终的模型拟合结果和预测结果。NDGM-ARIMA组合模型在血糖体检指标数据集上的预测结果表明，组合模型在体检指标序列上的预测精度有所提高，保证了预测结果的有效性和准确性，从而可以利用预测结果有效地分析出个人主要健康体检指标的变化趋势，实现人们健康管理的目标。

但是，本文模型存在一定的局限性。首先，本文研究数据集为等时距的近似非齐次指数序列，然而，实际应用中存在大量的非等间距的近似非齐次指数序列，容易导致因数据序列类型不符合预测模型而出现较大的建模误差。因此，如何进一步拓展灰色预测模型的适用范围将成为未来的研究方向。其次，本文组合模型中仅使用了方差倒数法求解各预测模型权重，但是单一赋权的方式可能存在较大的权重求解误差，因此在对多种赋权方法研究的基础上，是否可通过将两种及以上赋权方法结合起来进行求权，从而提高预测模型建模精度，同样是本文进一步的研究方向。

**参考文献：**


[1] 周伟杰, 张宏如, 党耀国, 等. 新息优先累加灰色离散模型的构建及应用 [J]. 中国管理科学, 2017, 25(8): 140–148.

[2] 曾波, 刘思峰, 曲学鑫. 一种强兼容性的灰色通用预测模型及其性质研究 [J]. 中国管理科学, 2017, 25(5): 150–156.

[3] 陆冬磊, 吴春峰, 段胜刚, 等. 应用灰色模型GM(1, 1)预测上海市副溶血性弧菌引起的食源性疾病发病率 [J]. 环境与职业医学, 2015, 32(8): 728–730.

[4] 曾波, 刘思峰, 白云, 等. 基于灰色系统建模技术的人体疾病早期预测预警研究 [J]. 中国管理科学, 2020, 28(1): 144–152.

[5] 华来庆, 申广荣, 熊林平, 等. ARIMA模型在黄瓜霜霉病疾病指数时间序列建模中的应用研究 [J]. 第二军医大学学报, 2006, 27(7): 729–732.

[6] VALIPOUR M, BANIHABIB M E, BEHBAHANI S M R. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir[J]. Journal of Hydrology, 2013, 476: 433–441.

[7] 刘琼, 杨建华. 隐马尔科夫模型在乙肝发病预测中的应用 [J]. 数学的实践与认识, 2017, 47(19): 203–210.

[8] 聂雄, 陈华, 伍思霖. 基于灰度共生矩阵和BP神经网络的乳腺肿瘤识别 [J]. 电子技术应用, 2019, 45(7): 97–101,116.

[9] 王振飞, 陈金磊, 郑志蕴, 等. 面向心血管疾病的自适应模块化神经网络预测模型 [J]. 小型微型计算机系统, 2019, 40(1): 232–235.

[10] ASILTÜRK İ, ÇUNKAŞ M. Modeling and prediction of surface roughness in turning operations using artificial neural network and multiple regression method[J]. Expert Systems With Applications, 2011, 38(5): 5826–5832.

[11] MCCLELLAND G H, IRWIN J R, DISATNIK D, et al. Multicollinearity is a red herring in the search for moderator variables: a guide to interpreting moderated multiple regression models and a critique of Iacobucci, Schneider, Popovich, and Bakamitsos (2016)[J]. Behavior Research Methods, 2017, 49(1): 394–402.

[12] SHIOTA S, OKAMOTO Y, OKADA G, et al. The neural correlates of the metacognitive function of other perspective: a multiple regression analysis study[J]. Neuroreport, 2017, 28(11): 671–676.

[13] YOO H Y, LEE J H, KIM D S, et al. Enhancement of glucose yield from canola agricultural residue by alkali pretreatment based on multi-regression models[J]. Journal of Industrial and Engineering Chemistry, 2017, 51: 303–311.

[14] 王永斌, 李向文, 柴峰, 等. 采用灰色–广义回归神经网络组合模型预测我国尘肺病发病人数的方法探讨 [J]. 环境与职业医学, 2016, 33(10): 984–987.

[15] 严薇荣, 徐勇, 杨小兵, 等. 基于ARIMA-GRNN组合模型的传染病发病率预测 [J]. 中国卫生统计, 2008, 25(1): 82–83.

[16] 时冬青, 宋文华, 张桂钏, 等. 基于灰色GM(1, 1)–马尔科夫模型的职业病预测研究 [J]. 中国安全生产科学技术, 2017, 13(4): 176–180.

[17] 范文俊, 刘静怡, 史菲, 等. 新型炎性指标对冠状动脉疾病的诊断预测价值 [J]. 医学研究杂志, 2021, 50(1): 80–85.

[18] 章玫红. 血清Hcy水平在心脑血管类疾病风险预测中的价值分析 [J]. 中国社区医师, 2021, 37(16): 103–104.

[19] 刘丹, 赵森, 颜志良, 等. 基于堆叠自动编码器的miRNA-疾病关联预测方法 [J]. 计算机科学, 2021, 48(10): 114–120.

[20] 苗立志, 白瑞思蒙, 刘成良, 等. 面向非平衡数据的癌症患者生存预测分析 [J]. 计算机工程, 2021, 47(12): 316–320.

[21] 谢乃明, 刘思峰. 一类离散灰色模型及其预测效果研究 [J]. 系统工程学报, 2006, 21(5): 520–523.


（编辑：丁红艺）