分析模型文档

1. 概述

本文档介绍分析模型的基本理论、实现方法、接口调用方法和相关算例。分析模型包括：

（1）常规数学模型：包括统计模型、灰色系统模型、时间序列模型、卡尔曼滤波模型

（2）智能算法模型：包括神经网络模型、支持向量机模型、极限学习机模型、模糊预测模型

（3）组合模型：包括统计-BP-SVM 组合模型，GM-BP-SVM 组合模型、趋势变化和周期波动组合模型、EEMD-SVM-ARMA 组合模型。

为了配合上述模型，还设计了数据预处理模块，下面将逐一进行介绍。分析模型采用 Python 语言开发，采用 WebAPI 方式调用，数据通过 JSON 格式定义，下面也会对输入格式进行说明。

2. 数据预处理模块

2.1 功能概述

数据预处理模块是所有分析模型的前置模块，负责接收用户传递的数据，进行加工处理后传递给后续分析模型。本模块采用 WebAPI 接口，数据通过 JSON 方式传递。

本模块的核心功能是因子处理。坝工理论和经验表明，大坝各种效应量主要受上下游水位、温度、时间效应等原因量的影响，但是这种影响不一定是线性的，以重力坝为例，大坝的变形往往与水位 H、H^2、H^3 具有相关性，在时效方面可能与时间 t、$\ln(1+t)$等具有相关性，因此如果直接用原始监测数据（H、t）作为原因量进行建模，则难以表达以上二次、三次和对数等关系，无法取得良好的建模效果。为解决上述问题，通常将效应量分解为水位（水压）、温度、时效等分量，每种分量均有一系列因子构成，例如：

（1）水压分量常考虑水深 H 的一次式~五次式作为因子，相应的表达式为：$\sum_{i=1}^{3\sim5} a_{1i} H^i$

（2）温度分量常考虑采用监测前 i 天的气温和水温的均值（T_i）或监测前 i 天的气温和水温与年平均温度的差值作为因子，相应的表达式为：$\sum_{i=1}^{m_2} b_{2i} T_i$

（3）时效分量考虑如下几种函数类型（其中 t 为时间）：

多项式：$\sum_{i=1}^{m} c_i t^i$

指数函数：$1-e^{-c_1 t}$

对数函数：$\ln(1+t)$

双曲函数：$\frac{\xi_1 t}{\xi_2 + t}$

三角函数：$\sum_{i=1}^{2}\left(c_{1i}\sin\frac{2\pi i t}{365} + c_{2i}\cos\frac{2\pi i t}{365}\right)$

不同的大坝、不同的效应量所适用的分量和因子表达式都有所不同，需要针对实际情况进行选取和设计，因此分量和因子的设置是一个需要经验和技巧的过程，具有很大的灵活性。针对上述问题，设计了专门的基于表达式解析的因子处理程序，通过数学表达式对各因子的数据处理方式进行

定义，程序会对表达式进行解析并执行相关处理预处理操作，具有很强的可定制性和通用性。

本模块除了提供因子处理功能外，还具备基本的重复项剔除等功能，确保为后续分析模型的提供有效的数据。

2.2 输入格式

接口数据参考如下范例，具体参数含义见表 2.2-1。


```
{
"Data" :  [ ["2000/01/01 00:00:00", 27.32, 17.1, 0.96, 0.2, 0.51, 0.6878],
            ["2000/01/02 00:00:00", 27.45, 15.2, 0.23, 0.93, 0.91, 0.6948],
            ["2000/01/03 00:00:00", 27.42, 14.2, 0.53, 0.22, 0.03, 0.7019], ... ]
"Col":    [ {"Item":"Time", "Type":"Time"},
            {"Item":"H1", "Type":"Head_Up"},
            {"Item":"T1", "Type":"Temp_Air"},
            {"Item":"J1", "Type":"Crack"},
            {"Item":"J2", "Type":"Crack"},
            {"Item":"J3", "Type":"Crack"},
            {"Item":"U5", "Type":"Disp"}, ...]
"Factor": [
{"Component":"Head", "ItemType":"Head_Up", "Expression":"None", "MaxOrder": 3},
{"Component":"Head", "ItemType":"Head_Up", "Expression":"Average(x,30)", "MaxOrder": 3},
{"Component":"Head", "ItemType":"Head_Up", "Expression":"x-Average(x,30)", "MaxOrder": 3},
{"Component":"Temp", "ItemType":"Temp_Air", "Expression":"Average(x,10)", "MaxOrder": 3},
{"Component":"Time", "ItemType":"Time", "Expression":"1-exp(-x)", "MaxOrder": 1},
{"Component":"Time", "ItemType":"Time", "Expression":"ln(1+x)", "MaxOrder": 1},
{"Component":"Time", "ItemType":"Time", "Expression":"sin(2*pi*x/365)", "MaxOrder": 1},
{"Component":"Time", "ItemType":"Time", "Expression":"cos(2*pi*x/365)", "MaxOrder": 1}, ...]
"Setting":{ "BaseTime": "2016/1/1 18:00:00",
            "Method": "Multiple"}
}
```


表 2.2-1 数据预处理模块接口参数列表

名称		类型	说明	备注
xData		float[][] 二维数组	原因量数据，第一列为时间，用字符串表示，时间格式为 yyyy/mm/dd HH:MM:SS；后面为具体监测数据。	可以看成矩阵，每行表示同一时间不同测点的数据，每列表示同一个测点各时间的数据。 考虑到兼容性，用字符串表示时间更为可靠。
xCol		object[] 对象数组	描述 xData 每列数据的含义	
xCol 对象	Item	string	列标记，用于唯一标记测点，不能重复	
	Type	string	列类型，用于表示数据类型，例如：Time, Head_Up, Head_Down, Temp_Air, Crack 等	分别表示时间、上游水位、下游水位、气温、裂缝等分量
yData		float[][]	效应量量数据，第一列为时间，用	其定义方法与 xData 和

		二维数组	字符串表示，时间格式为 yyyy/mm/dd HH:MM:SS；后面为具体监测数据。	xCol 相同
yCol		object[] 对象数组	描述 yData 每列数据的含义	
Factor		object[] 对象数组	描述分析因子设置	
Factor 对象	Component	string	因子所属分量，有效值为：Head, Temp, Time, Crack	分别表示水压、温度、时效和裂缝四类因子
	ItemType	string	数据类型，例如：Time, Head_Up, Temp_Air, Temp_Water, Temp_Dam, Crack 等	与 xCol.Type 一致
	Expression	string	对数据的进一步操作表达式，如果直接用原始数据，不做任何操作，则设为 None，否则通过数学表达式形式描述，其中数据用 x 表示，例如：ln(1+x), Average(x,10), AverageRange(x,i,j)等，支持常用的基本数学函数。	Average(x,i)为特殊函数，表示前 i 天的平均值；AverageRange(x,i,j)biao'shi 前 i 天~前 j 天的平均值。
	MaxOrder	int	分量的最高多项式阶次，例如对上游水位 H_{up} 选取 5，表示 $\sum_{i=1}^{5} c_i H_{up}^i$	该值不小于 1
Setting		object	模型相关参数	
Setting 对象	BaseTime	string	基准时间，格式为 yyyy/mm/dd HH:MM:SS	用于计算经过的天数等
	选项参数（取决于具体模型）	取决于具体模型	针对具体模型的选项	针对具体模型有不同的选项参数，详见各模型部分说明。
FileName		string	模型文件名	用于存储和读取模型

上述输入格式主要针对建模（模型训练）阶段，当模型建立完成后使用模型进行预测时，不需要再提供 Factor 和 Setting 参数（模型文件内部会有保存）。

2.3 处理过程

因子处理过程中，针对每项因子设置，程序的执行流程如下：

①根据 ItemType，在输入数据中找到对应的列，存入变量 x；

②根据 Expression 的表达式，由内置的表达式解析模块进行解析，按照表达式对 x 进行处理，得到处理后的 x；

③根据 MaxOrder 对处理后的 x 进行扩充，如果 MaxOrder=3，那么在 x 基础上，进一步计算 x^2 和 x^3；

④存储当前因子的处理结果，转入下一因子，重复上述过程。

下面通过实例介绍因子处理的计算过程，假设有如下表所示的原因量数据：

表 2.3-1 因子处理前的原始原因量数据

因子处理方面的典型输入信息如下（JSON 格式）：

"Factor": [

{"Component": "Head", "ItemType": "Head_Up", "Expression": "None", "MaxOrder": 3},

{"Component": "Head", "ItemType": "Head_Up", "Expression": "x-Average(x,5)", "MaxOrder": 3},

{"Component": "Temp", "ItemType": "Temp_Air", "Expression": "None", "MaxOrder": 1},

{"Component": "Temp", "ItemType": "Temp_Air", "Expression": "AverageRange(x,3,5)", "MaxOrder": 1},

{"Component": "Time", "ItemType": "Time", "Expression": "ln(1+x)", "MaxOrder": 1}　]

以上输入信息用表格展示，如下表所示：

表 2.3-2 因子处理输入信息

Component（表示分量类型）	**ItemType（表示同一分量下的具体测点类型）**	**Expression（因子计算表达式）**	**MaxOrder（因子最高多项式阶数）**	**备注**
Head（水压分量）	Head_Up（上游水位）	None（不做任何处理）	3（表示 1，2，3 阶）	即水位 H、H^2、H^3
Head（水压分量）	Head_Up（上游水位）	x-Average(x,5)（当前水位与前 5 天平均水位之差）	3	Average(x,k)为特殊函数，表示前 k 天平均值
Temp（温度分量）	Temp_Air（气温）	None	1	即气温 T
Temp（温度分量）	Temp_Air（气温）	AverageRange(x,3,5)	1	AverageRange(x,k1,k2)为特殊函数，表示前 k1~k2 天平均值
Time（时效分量）	Time（时间，以天为单位）	ln(1+x)	1	支持 ln,exp,sin,cos 等常用数学表达式

除因子设置信息外，模型还需要一些选项，典型输入如下：

"Setting":{"BaseTime" : "2011-01-11 00:00:00", "Method":"Multiple", "Intercept":"Yes"}

其中，BaseTime 表示建模基准时间，在计算时效分量时会将当前时间换算为距离 BaseTime 的

天数（支持小数）；Method 和 Intercept 是用于统计模型的选项，决定了回归建模方法和截距的处理，对于不同的模型有不同的选项，详见具体模型部分。

对于上述输入信息，程序充分考虑了通用性和可扩充性，有如下几个特点：

（1）Component（分量类型名）中采用的 Head、Temp、Time 等并非程序限定的表达，可以任意用其他英文单词代替，例如用 WaterLevel 替代 Head，程序则会返回名为 WaterLevel 的分量结果，也可以增加例如 Rain 表示降雨量分量，完全由前端页面进行配置，可以自由扩充。

（2）ItemType（测点类型名）中采用的 Head_Up、Temp_Air、Time 同样并未限定，只要能够在表 1 的原因量中存在对应名字的列，程序就可以提取相应的数据，因此并不局限于有限的几种测点类型，可以自由扩充。

（3）Expression（分量表达式）中支持各种常用数学函数和运算符，以及两个特殊函数 Average 和 AverageRange（含义见表 2.2 备注），程序会对表达式进行解析，因此具有很强的灵活性和扩展性。注意表达式中原因量统一用 x 表示。

（4）MaxOrder 表示最高多项式阶数，因此 MaxOrder=3 时即包含了一次方、二次方、三次方，这样处理可以在程序中减少重复计算，提高效率；如果有特殊需要，例如只考虑一次方和三次方，那么也可以通过 Expression=x，MaxOrder=1 和 Expression=x^3，MaxOrder=1 来实现（^表示乘方）。

（5）对于时间，因子处理时统一以天为单位，如果要考虑以月或年为单位，可以在表达式 Expression 中进行转换，例如 ln(1+x/30)、ln(1+x/365)等。

根据上述流程，利用表 2.3-2 中的因子设置对表 2.3-1 进行处理，原始原因量有 Time、Head_Up、Temp_Air 共 3 项，因子处理后将扩充为表 2.3-3 所示 9 项原因量，程序统一用 x1, x2, x3...进行表示，因子处理后原因量数据如表 2.3-4 所示。

表 2.3-3 因子处理后原因量项目

表 2.3-4 因子处理后原因量数据

对比表 2.3-1 和表 2.3-4 可以看出，因子处理过程将原始原因量进行了扩充，利用扩充后原因量进行建模，有望取得更为理想的效果。

2.4 其他说明

为给建模提供更多参考信息，因子处理完毕后，自动进行原因量和效应量的相关性分析，为建模提供参考。相关性分析会计算相关系数、偏相关系数和半偏相关系数，结果通过日志的形式返回，典型结果如图 2.4-1 所示。

原因量	相关系数	偏相关系数	半偏相关系数
x1	-0.3551183043478155	0.0001931259844998234	0.21240463375244853
x2	-0.35655800113375585	0.03285130430494349	0.023801908715762335
x3	-0.35764227774842783	-0.031446624079089415	-0.022783139050655567
x4	-0.34147285644981545	0.0229560306238504	0.05882415037254338
x5	-0.34284930886108017	-0.026261616467847297	-0.01902374390410886
x6	-0.3438854083090959	0.02363582166949708	0.017120512977852893
x7	-0.06288829161077873	-0.00038800224470850886	0.1853957198119175
x8	-0.016552749018214383	-0.09433277966208406	-0.06861644929003002
x9	-0.30436390126745944	-0.07383742705883686	-0.053615251501775506
x10	-0.32863548984570784	-0.05801963429930078	-0.04208544090961906
x11	-0.33073452893176886	0.12155319275348772	0.08867953054645661
x12	-0.32894378168181526	0.003996415418259783	0.002893999308559436
x13	-0.48972282423724367	0.03601057964352948	0.02606438771983789
x14	-0.29246647046082125	-0.04743293588812598	-0.034348224001735816
x15	-0.4768219011296536	0.04742407821204868	0.034341831254766746
x16	-0.48374980056825917	-0.042110607312347344	-0.030486850940413254
x17	0.25238748656468035	0.147600517737782	0.10806762337716125
x18	0.23283485445378713	0.09564003478144076	0.06957605243567272
x19	-0.0629175572795484	0.034145113556586726	0.024740390825894285
x20	-0.007035723074325824	0.006832634628612824	0.004947923117454874

图 2.4-1 相关性分析结果示例

因子设置需要一定的经验和判断，不能随意的设置，如果设置的因子不合理，可能导致处理后原因量中出现常数列或者高度相关的两列，这会造成回归分析中的“多重共线性”问题，使得模型估计失真或难以估计准确。

如果因子处理后的原因量数据的时间点跟效应量不对应，程序会使用线性插值方法进行处理，确保原因量和效应量的时间对应。

3. 统计模型

3.1 功能概述

统计模型基于线性回归理论，是最经典和最常用的大坝分析模型。针对各类统计模型，包括变形统计模型、裂缝开合度统计模型、应力统计模型、渗流统计模型等，分析其共同点，编制了通用的统计模型模块。

本模块主要提供模型训练（建模）和模型预测两方面功能：

（1）建模：根据用户选择的因子设置和某一时间段的实测资料，由平台生成特定格式的输入数据（格式要求见数据预处理模块）并提交给本模块，本模块采用线性回归方法确定模型中各项因子的系数，建立回归模型并输出相关参数和结果；

（2）预测：根据用户选择的模型和某一时间段的实测资料，由平台生成特定格式的输入数据（格式要求见数据预处理模块）并提交给本模块，本模块进行推算并输出预测结果给平台，由平台进一步可视化。

本模块提供了多元线性回归和逐步回归两种求解方法，支持有常数项和无常数项两种方式；除输出整体拟合或预测数据外，还能够输出水压、温度、时效等分量数据；能够输出实测值和计算值的 RMSE（均方根误差）、R^2（确定系数）、R（相关系数）等参数用于评估结果。下面将对该模型进行详细说明。

3.2 基本理论

3.1.1 基本概念

在建模时，首先有原因量和效应量两个基本概念：

（1）效应量：也称为因变量，即建模的目标。例如想要对大坝某测点水平位移进行建模，则该水平位移即为效应量，此外渗流、应力、裂缝等监测量均可作为效应量。

（2）原因量：也称为自变量，即可能引起效应量变化的因素，例如水位、温度、时间、降雨量等等。

大坝的效应量按成因可分为：水压分量、温度分量、时效分量、降雨分量、裂缝分量（某些大坝在下游面有较大范围水平裂缝时需要考虑）等。基于大坝和坝基的力学和结构理论分析，常用的分量因子及其表达式如下：

（1）水压分量，可考虑如下因子：

①水深 H 的一次式~五次式，相应的表达式为：

$$\sum_{i=1}^{3\sim5} a_{1i} H^i$$

②测值前的月平均水深 H_1 的一次式~三次式，相应的表达式为：

$$\sum_{i=1}^{3} a_{2i} H_1^i$$

③监测时水深与监测前 j 天平均水深之差 $\Delta\bar{H}_j$ 的一次式~二次式，相应的表达式为：

$$\alpha_f \Delta\bar{H}_j + \alpha_b \left(\Delta\bar{H}_j\right)^2$$

（2）温度分量，可考虑如下因子：

①完全无温度资料时，采用多周期谐波因子，相应的表达式为：

$$\sum_{i=1}^{1(2)} \left(b_{1i} \sin\frac{2\pi it}{365} + b_{2i} \cos\frac{2\pi it}{365} \right)$$

其中，t 为监测日至始测日的累计天数。

②有气温或水温资料 T 时，采用监测前 i 天的气温和水温的均值（T_i）或监测前 i 天的气温和水温与年平均温度的差值作为因子，相应的表达式为：

$$\sum_{i=1}^{m_2} b_{2i} T_i$$

③有混凝土温度资料 T 时，采用温度计测值或等效温度作为因子，相应的表达式为：

$$\sum_{i=1}^{m} b_{3i} T_i \text{ 或 } \sum_{i=1}^{m} b_{3i} \overline{T}_i + \sum_{i=1}^{m} b_{4i} \beta_i$$

其中，m 为温度计或等效温度数目。

（3）时效分量，可考虑如下几种函数类型（其中 θ 为时间）：

①指数函数：$C\left[1-e^{-c_i\theta}\right]$

②双曲函数：$\dfrac{\xi_1\theta}{\xi_2+\theta}$

③多项式：$\sum_{i=1}^{m} c_i \theta^i$

④对数函数：$c\ln(1+\theta)$

⑤指数函数（或对数函数）的附加周期项：$\sum_{i=1}^{2} \left(c_{1i} \sin\frac{2\pi it}{365} + c_{2i} \cos\frac{2\pi it}{365} \right)$

（4）降雨分量，可考虑为降雨量的多项式函数，相应的表达式为：$\sum_{i=1}^{m_3} d_i R^i$，其中 m_3 为考虑的多项式阶次。

（5）裂缝分量

可选用测缝计的开合度 J 的测值作为因子，即：$\sum_{i=1}^{m_4} f_i J_i$，其中 m_{34} 为测缝计个数。

统计模型除以下各种分量和因子外，通常还补充一个常数项。除上述几种分量类型外，还可以具体的坝型和监测数据，设计合适的分量和因子。在设计好后，通过数据预处理模块进行因子处理，获得用于建模的数据。

3.2.2 多元线性回归理论

统计模型建模即利用多元线性回归理论，建立原因量和效应量之间的线性回归模型，该模型为多项式形式表达，其一般形式为：

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m \tag{3.1}$$

式中，y 为效应量，$x_1, x_2, ..., x_m$ 为原因量，系数 $a_0, a_1, ..., a_m$ 一组未知的常数，称为回归系数，其中 a_0 称为常数项（截距）。回归的目的就是希望通过对 y 和 x 的一系列观测值来估计 a_i (i=0,1,2,...,m)的值。一般来说，多元线性回归模型包含常数项，有特殊要求时也可能不包含常数项，此时称为过原点（无截距）的线性回归。

假设效应量 y 和因子处理后原因量 x 共 n 次观测数据，分别记为 $y(k)$和 $x_i(k)$，i=1,2,3,...,m；k=1,2,3,...,n，根据式(3.1)，可用 n 个线性方程组表示上述观测关系：

$$\begin{aligned} y(1) &= a_0 + a_1 x_1(1) + a_2 x_2(1) + \cdots + a_m x_m(1) \\ y(2) &= a_0 + a_1 x_1(2) + a_2 x_2(2) + \cdots + a_m x_m(2) \\ &\cdots \\ y(n) &= a_0 + a_1 x_1(n) + a_2 x_2(n) + \cdots + a_m x_m(n) \end{aligned} \tag{3.2}$$

上式可以用矩阵形式表示：

$$\begin{Bmatrix} y(1) \\ y(2) \\ \vdots \\ y(n) \end{Bmatrix} = \begin{bmatrix} 1 & x_1(1) & x_2(1) & \cdots & x_m(1) \\ 1 & x_1(2) & x_2(2) & \cdots & x_m(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1(n) & x_2(n) & \cdots & x_m(n) \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{Bmatrix} \tag{3.3}$$

可简记为：

$$\boldsymbol{y = Xa} \tag{3.4}$$

如果原因量和效应量关系精确的符合上式，且观测数据准确，无随机干扰项，模型的回归系数 a 就可仅由 n=m 组观测数据唯一确定，即

$$\boldsymbol{a = X^{-1} y} \tag{3.5}$$

由于实际原因量和效应量的关系并不完全满足式(3.4)，同时观测值不可避免的存在误差，并受到各种干扰，导致 n=m 组观测数据唯一确定的回归系数向量值并不可靠，此时模型的回归系数向量 $\boldsymbol{a}$ 需要用更多的观测数据来进行估计，即 $n > m$。当 $n > m$ 时，由于线性方程组的个数大于回归系数的个数，不可能找到一组精确的解来满足所有 n 个方程。此时只能从许多近似解中找出一组残差平方和最小的估计解，即最小二乘解答。由于模型和观测值的误差以及噪声干扰，式（3.4）可改写为：

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{a}+\boldsymbol{e} \tag{3.6}$$

其中，$\boldsymbol{e}$ 为误差向量。上述方程组的“最小二乘”估计就是要找到一个回归系数向量，使观测数据方程组中的误差满足下面函数的要求：

$$\min J=\min\sum_{k=1}^{n}e^{2}\left(k\right)=\min\boldsymbol{e}^{T}\boldsymbol{e} \tag{3.7}$$

上式中 J 可改写为：

$$J=\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{a}\right)^{T}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{a}\right)=\boldsymbol{y}^{T}\boldsymbol{y}-\boldsymbol{a}^{T}\boldsymbol{X}^{T}\boldsymbol{y}^{T}-\boldsymbol{y}^{T}\boldsymbol{X}\boldsymbol{a}+\boldsymbol{a}^{T}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{a} \tag{3.8}$$

求上式的极值：

$$\left.\frac{\partial J}{\partial\boldsymbol{a}}\right|_{\boldsymbol{a}=\hat{\boldsymbol{a}}}=-2\boldsymbol{X}^{T}\boldsymbol{y}+2\boldsymbol{X}^{T}\boldsymbol{X}\hat{\boldsymbol{a}}=0 \tag{3.9}$$

由此可得回归系数向量的最小二乘估计：

$$\hat{\boldsymbol{a}}=\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{y} \tag{3.10}$$

在特殊情况下不考虑常数项时，式(3.3)变为：

$$\begin{Bmatrix}y(1)\\y(2)\\\vdots\\y(n)\end{Bmatrix}=\begin{bmatrix}x_1(1)&x_2(1)&\cdots&x_m(1)\\x_1(2)&x_2(2)&\cdots&x_m(2)\\\vdots&\vdots&\vdots&\vdots\\x_1(n)&x_2(n)&\cdots&x_m(n)\end{bmatrix}\begin{Bmatrix}a_1\\a_2\\\vdots\\a_m\end{Bmatrix} \tag{3.11}$$

此时同样可以按照式(3.10)计算最小二乘估计。

3.3.3 逐步回归理论

建立多元线性回归模型时，为了保证回归模型具有优良的解释能力和预测效果，原因量应符合一定原则：

（1）原因量对效应量必须有显著的影响，并呈密切的线性相关；

（2）原因量之间应具有一定的互斥性，即原因量之间不应当高度线性相关。

上述条件，特别时第 2 条，是线性回归的基本假设。然而在实际回归建模时，可能出现原因量对效应量并无显著影响，或者不同原因量之间存在较强的相关性，此时会出现**多重共线性**问题，使得模型估计失真或难以估计准确，建立的回归模型也失去效果。

此外，在回归建模时也经常会遇到自变量数目过大的情况，比如：影响大坝运行安全的预报因子按其成因可以分为三个部分，水压分量、温度分量和时效分量，此处各个分量又由许多分项组成，这就使得所给的分项较多，若将所有分项直接引入方程，可能会使系数矩阵蜕化和病态，进而影响回归方程的精度，甚至于无法求解。

针对上述问题，逐步回归方法是一个较好的解决策略。逐步回归方法即根据对效应量贡献的大小，有选择的将因子依次放入回归方程，进而建立最佳回归方程，其实质就是对多元线性回归的优化，它是在多元线性回归分析的基础上派生出的一种研究和建立优化多元回归方程的算法技巧。

逐步回归的基本思想是基本思想：逐个引入原因量，每次引入影响最显著的原因量，并对方程中的老变量逐个进行检验，把变得不显著的变量逐个从方程中剔除，最终的回归方程中既不漏掉影响显著的变量，又不包含影响不显著的变量。

本程序逐步回归方法的算法流程如下：

（1）首先逐个比较各原因量对回归方程的显著程度，即依次将原因量单独放入回归方程，进行

F 检验，挑选 P 值最小的原因量引入回归方程，即对方程贡献最大的原因量（F 检验相关算法较为复杂，程序调用专门的模块进行计算）；

（2）进入：在方程中已有原因量基础上，依次将未选择的原因量逐个单独放入回归方程，每次放入后进行 F 检验，然后挑选 P 值最小的原因量，即对方程贡献最大的原因量，如果 P 值小于进入阈值（0.05），则该原因量加入回归方程；

（3）退出：在回归方程中有新的原因量进入后，还需要考虑新增加原因量对旧原因量的影响，可能因为新加入的原因量导致旧的原因量不再显著。因此，针对新加入原因量后的回归模型再次执行 F 检验，统计每个原因量的 P 值，挑选 P 值最大的原因量，，即对方程贡献最小的原因量，如果 P 值大于退出阈值（0.1），则该原因量退出回归方程；

（4）反复执行（2）和（3）步，直至没有原因量进入和退出，得到回归模型。

当使用逐步回归时，程序会记录每个原因量的选择情况，没有选择的原因量将不进入回归方程中，也不参与后续计算。

3.3 模型接口

本模型提供训练和预测两个接口，为 WebAPI 形式，通过 POST 发送 JSON 格式请求到对应地址进行调用，具体地址如下：

训练：/AnalysisModel/Stats/Train

预测：/AnalysisModel/Stats/Predict

接口输入数据格式由数据预处理模块定义，详见 2.2 节。原因量（xData，xCol）、效应量（yData，yCol）、因子设置（Factor）和模型选项（Setting）均采用统一的模式。

调用模型训练接口时，需要提供原因量、效应量、因子设置和模型选项四个方面的数据和参数，调用模型预测接口时，有如下要点：

（1）预测用的原因量必须跟训练用的原因量完全对应，例如训练时采用了气温作为原因量，那么预测时必须提供气温数据，否则程序无法执行；如果预测时额外提供了训练时没有用到的数据，则程序会自行忽略；

（2）统计模型训练时，程序会保存相应的因子设置和模型选项，因此预测时不需要提供因子设置和模型选项，程序会根据保存的信息自动执行因子处理，从原始原因量获得处理后原因量，该过程与训练时的因子处理过程一致；

（3）在获得处理后原因量后，程序将其组成矩阵 $\boldsymbol{X}$，并根据训练模型是否有常数项，将 $\boldsymbol{X}$ 代入式(3.3)或式(3.11)进行计算，即可得到 y，即预测的效应量。

针对本模型，选项（Setting）中除了需要指定基准时间（BaseTime）外，还有如下特有选项：

表 3.3-1 统计模型选项

参数名	含义	取值范围	默认值
Method	回归建模方法	'Multiple'（多元线性回归） 'Stepwise'（逐步回归	'Multiple'
Intercept	是否考虑常数项（截距）	'Yes'（有常数项/截距） 'No'（无常数项）	'Yes'

如果不人工指定模型选项，默认情况下 Method=Multiple，Intercept=Yes，此时为考虑常数项的多元线性回归。程序将所有的原因量数据组成矩阵 $\boldsymbol{X}$，将所有效应量数据组成向量 $\boldsymbol{y}$，为考虑常数项的影响，程序会按式(3.3)所示，在 $\boldsymbol{X}$ 左边插入一列值为 1 的向量，然后按照式(3.10)计算回归系数估

计值；如果指定 Intercept=No，则程序会按照无常数项处理，即不会在 ***X*** 左边插入一列值为 1 的向量，而是直接利用原因量矩阵 ***X***，按照式(3.10)计算回归系数估计值。如果指定 Method=Stepwise，程序会按照 3.2.3 节给出的算法进行逐步回归。

3.4 模型输出

3.4.1 输出格式

返回结果为 JSON 格式，典型结果如下，上述结果中具体参数的含义见表 3.4-1。


```
{
    "StatusCode": 200,
    "StatusMessage": "成功！",
    "ModelFile":["D:\test.bin"],
    "Time": ["2013-10-14 00:00:00", "2013-10-19 00:00:00", "2013-10-30 00:00:00",...],
    "yReal": [5.0,5.32,5.78,5.92,5.99,6.43,6.58,6.6,...],
    "yCalc": :[4.7995475,5.6393201941,5.6608651355,6.064281,....],
    "yComponent": {"Head":[6.31,5.26,3.23,...], "Time":[4.21,5.33,3.86,...],...}
    "Factor": [{"Component":"Const","Expression":"","Item":"","ItemType":""}, {"Component":
            "Head", Expression":"x-75","Item":"Head_Up","ItemType":"Head_Up"},...],
    "Evaluate": {"R":0.98,"R2":0.97,"R2_adj":0.97,...],
    "Formula": " y = 46.1416+0.314735*x1-0.0013346*x2+0.0826778*x3"
    "Summary": "....................."
}
```


表 3.4-1 统计模型输出信息

参数	类型	含义
yReal	数组	原始效应量数据，跟输入值相同
yCalc	数组（跟 yCalc 等长）	模型计算出的效应量数据
yComponent	对象	各分量数据
Summary	字符串	建模过程输出信息，包含护输出输入原因量和处理后原因量，总体精度，方差分析表，回归系数表等。
Factor	对象	处理后各原因量 备用
xProdcessed	数组	处理后原因量数据 供日后检验 备用
Evaluate	对象	包含大量模型精度检验相关数据，详细内容见下表
Formula	字符串	回归公式，与 Evaluate 中的 expr 项相同
ModelFile	字符串	模型文件路径
StatusCode	整数	状态号，成功为 200，有错误为 500
StatusMessage	字符串	如果发生错误，存储错误提示信息，详细信息在 Summary 中

3.4.2 Evalaute 项含义

表 3.4-1 中，Evaluate 中的各项含义如下：

表 3.4-2 Evaluate 中各项含义

参数	类型	含义
variable	数组	原因量名称
param	数组	非标准化系数
param_se	数组	系数标准误
param_t	数组	系数 pvalue

param_p	数组	系数显著性(P>|t|)
beta	数组	标准化系数
corr	数组	各原因量相关系数
pcorr	数组	各原因量偏相关系数
spcorr	数组	各原因量半偏相关系数
pR2	数组	各原因量偏决定系数
vif	数组	方差膨胀系数
eigenval	数组	特征值
R	实数	复相关系数
R2	实数	决定系数
R2_adj	实数	调整后决定系数
RMSE	实数	估计标准误差
dof_regress	实数	回归自由度
dof_residual	实数	残差自由度
dof_total	实数	总自由度
mse_regress	实数	回归均方和
mse_residual	实数	残差均方和
ssr	实数	回归平方和
sse	实数	残差平方和
sst	实数	总平方和
fvalue	实数	F 统计量
f_pvalue	实数	显著性
cond	实数	条件数
expr	字符串	回归方程，跟 Formula 相同

3.4.3 Summary 项含义

Summary 中会给出模型运行过程各详细信息，还会提供一些信息表格，如图 3.4-1 所示。

相关系数(R)	决定系数(R2)	调整后决定系数(R2-adj)	估计标准误差(std err)
0.983	0.965	0.963	0.01191

来源	平方和	自由度	均方和	F统计量	显著性P
回归	0.862	17	0.051	357.938	0.000
残差	0.031	218	0.000		
综合	0.893	235			

原因量	非标准化系数(B)	标准误	标准化系数(Beta)	t	显著性(P>|t|)	相关系数	偏相关系数	半偏相关系数	偏决定系数	方差膨胀系数(VIF)
Const	-1174.346	248.597		-4.724	0.000					
x15	0.082	0.001	0.943	70.219	0.000	0.955	0.977	0.936	0.956	1.015
x14	-0.360	0.064	-0.984	-5.674	0.000	-0.272	-0.350	-0.076	1.000	169.464
x18	-0.009	0.001	-0.102	-7.627	0.000	-0.085	-0.449	-0.102	0.204	1.011
x13	1177.504	249.186	0.820	4.725	0.000	-0.268	0.297	0.063	1.000	169.495
x17	-0.004	0.001	-0.048	-3.587	0.000	-0.052	-0.230	-0.048	0.053	1.006

图 3.4-1 Summary 中输出信息表示例

上述信息形成三张表：总体精度表，方差分析表(ANOVA)，回归系数表(Coefficients)，表中相关项对应表 3.4-2 中 Evalaute 中的项目，具体定义见表 3.4-2~表 3.4-4。

表 3.4-2 总体精度表对应 Evaluate 中项目

相关系数(R)	决定系数(R2)	调整后决定系数(R2-adj)	估计标准误差(std err)
R	R2	R2_adj	RMSE

表 3.4-3 方差分析表(ANOVA) 对应 Evaluate 中项目

来源	平方和	自由度	均方和	F 统计量	显著性 P
回归	ssr	dof_regress	mse_regress	fvalue	f_pvalue
残差	sse	dof_residual	mse_residual		
综合	sst	dof_total			

表 3.4-4 回归系数表(Coefficients) 对应 Evaluate 中项目

原因量	非标准化系数(B)	标准误	标准化系数(Beta)	t	显著性(P>|t|)	相关系数	偏相关系数	半偏相关系数	偏决定系数	方差膨胀系数(VIF)
variable	param	param_se	beta	param_t	param_p	corr	pcorr	spcorr	pR2	vif

3.4.4 回归方程

程序返回的回归方程格式如下：

y=-9.47607+ 0.16371*x1+ 4.10804e-05*x2+ 1.37275e-05*x3+ 0.294903*x4 -0.0960315*x5 -0.0322229*x6-0.616101*x7

其中 x1~x7 的含义需要参考返回的 Factor 项来了解，为增强可读性，回归方程中常用特定符号表示原因量，此时需要前台对回归方程进一步格式化。下面给出格式化的具体建议：变量通常用一个字母表示，用上标表示多少次方，下标表示补充信息。

（1）水位通常用 H 表示，例如 H， H^2， H^3

如果要区分上下游水位，通常通过下标进行区分，例如 $H^2_{上游}$， $H^2_{下游}$， $H^2_{水位差}$

如果要表示前 30 天平均水位，15~30 天平均水位，也通过下标区分，例如 H^2_{30}、 $H^2_{15\sim30}$

如果同时表示上下游和前 30 天平均水位，可以采用 $H^2_{上游30}$

（2）温度通常用 T 表示，例如 T， T^2， T^3

如果要表示前 30 天平均气温，15~30 天平均气温，通过下标区分，例如 T^2_{30}、 $T^2_{15\sim30}$

（3）时间通常用 t 或 θ 表示，例如 $\ln(1+t)$， $\ln(1+\theta)$

（4）降雨量可以用 R 表示，例如 R， R^2， R^3

其他的变量可以自己约定字母表达，没有通用的表达方式。

3.4.5 分量输出

模型除了可以输出效应量的整体计算结果外，还可以输出各个分量的计算结果。以表 2.3-1 所示的因子处理信息和表 2.3-2 所示的处理后原因量项目为例，介绍具体计算过程：

（1）在因子处理过程中，程序会标记各个分量对应的原因量，如下表所示：

表 3.4-5 各分量对应的原因量

分量类型	**对应的原因量**
Head	x1, x2, x3, x4, x5 ,x6
Temp	x7, x8
Time	x9

（2）计算每个分量类型引起的效应量，以水压分量 Head 为例，在式(3.3)的回归系数向量中屏蔽掉常数项、Temp 和 Time 的影响，即令 a_0, a_7, a_8, a_9 均为 0，然后将原因量矩阵 $\boldsymbol{X}$ 代入计算，如下式所示，即可得到对应的效应量分量值。如果要考虑温度分量 Temp，则在回归系数向量中屏蔽掉常数项、Head 和 Time 的影响，即令 $a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_9$ 均为 0，即可得到温度分量值。

$$\begin{Bmatrix} y(1) \\ y(2) \\ \vdots \\ y(n) \end{Bmatrix} = \begin{bmatrix} 1 & x_1(1) & x_2(1) & x_3(1) & x_4(1) & x_5(1) & x_6(1) & x_7(1) & x_8(1) & x_9(1) \\ 1 & x_1(2) & x_2(2) & x_3(2) & x_4(2) & x_5(2) & x_6(2) & x_7(2) & x_8(2) & x_9(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1(n) & x_2(n) & x_3(n) & x_4(n) & x_5(n) & x_6(n) & x_7(n) & x_8(n) & x_9(n) \end{bmatrix} \begin{Bmatrix} 0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ 0 \\ 0 \\ 0 \end{Bmatrix} \tag{3.12}$$

（3）如果建立的回归模型没有采用常数项，计算方法与有常数项时是一致的，计算时采用式(3.11)。

（4）计算出来的分量值是对效应量的分解，因此跟效应量具有相同的物理意义和单位，例如效应量是水平位移，则各分量的含义均为水平位移。

（5）当回归模型采用常数项时，各分量值相加后还需增加常数项后才能与效应量计算值一致，程序会将常数项单独输出作为一个特殊分量。

3.5 其他说明

在统计模型建模时，可能出现模型整体拟合较好，但分量值异常的现象，这主要是因为因子设置的不合理，简单解释如下：

（1）因子设置导致处理后原因量中存在高度线性相关项，引发多重共线性问题，例如使用当前水位和 3 天平均水位作为因子，当水位变化不大时，这样的两组数据很可能高度一致，这种问题会导致模型估计失真或难以估计准确，从而出现异常的分量值；

（2）时效因子的计算过程的单位选取不合理，当数据时间跨度较大时，默认以天为单位计算时效因子会导致不合理的值，此时需要对单位进行转换，例如采用 ln(1+*t*/30)转换为以月单位，则有望改善分量结果；

（3）因子设置与实际之间存在偏差，例如用大坝的水位作为因子来对位移进行建模，水位通常是海拔高度，而理论上对大坝位移有影响的是净水位（水位-坝底高程）或上下游水位差，因此考虑考虑对上游水位减去固定的值，或者用上下游水位差来计算水压分量，有望改善分量结果。

（4）如果一次性选用了太多的因子，会导致模型非常复杂，默认的多元回归方法会考虑所有原因量（即便这些原因量不显著甚至有害），导致结果不理想。此时，可以考虑选用逐步回归方法进行优化。

（5）没有考虑初始状态的影响。以位移测点为例，假设回归方程为：

$$\delta = a_0 + a_1 H + a_2 H^2 + a_3 H^3 + a_4 T + a_5 t$$

式中：H 为水位，T 为温度，t 为时间。设状态 I 为大坝无变形状态，对应的水位、温度、时间为 0，状态 II 为测点起测时间，对应的水位、温度和时间分别为 H_0、T_0、t_0，当前时刻的监测值是相对于状态 II 的变形值，而上式描述的却是相对于状态 I 的变形值，与监测值有差别，故回归分析时必然存在常数项，由于水位分量多项式阶数较高，可能会产生很大的常数项。如果采用下式的回归模型，扣除了起测时间时的水位、温度和时效基准值，则回归方程中的常数项就能显著减小

$$\delta = a_0 + a_1 (H - H_0) + a_2 (H - H_0)^2 + a_3 (H - H_0)^3 + a_4 (T - T_0) + a_5 (t - t_0)$$

综上，统计模型的分量是否合理是一个需要针对具体问题反复调试的过程。由于分析者选择的因子设置不合理而导致分量结果不合理，并不代表程序计算有误，即便换用其他统计分析软件进行人工分析，得到的结果依然是一样的。程序作为计算工具忠实执行分析者的指令，而能否建立优良的模型则依赖于分析者自身的理论、经验和技巧。

3.6 算例验证

3.6.1 算例说明

选用大坝 PL01XH01 测点 2013/10~2020/9 的监测数据，利用 2013~2019 年数据进行建模，然后预测 2020 年数据与实测数据进行比对，以验证模型的有效性。2013~2020 年环境量（原因量）和测点实测数据（效应量）的过程线如下图所示。

图 3.6-1 环境量和测点实测数据过程线

3.6.2 因子设置

分析原因量和效应量的关系，选择因子设置如下

Component（分量类型）	**ItemType（测点类型）**	**Expression（因子计算表达式）**	**MaxOrder（最高多项式阶数）**	**备注**
Head（水位分量）	Head_Up（上游水位）	x-75	2	
Head（水位分量）	Head_Down（下游水位）	x-75	2	
Temp（温度分量）	Temp_Air（气温）	x	1	
Temp（温度分量）	Temp_Air（气温）	Average(x,30)	1	30 天平均气温
Time（时效分量）	Time（时间）	ln(1+x/365)	1	x/365 转换时间单位为年

3.6.3 模型选项

选用多元线性回归，考虑常数项，即：

选项	含义	取值	备注
Method	回归建模方法	Multiple	多元线性回归
Intercept	是否考虑常数项	Yes	考虑常数项

3.6.4 总体结果

利用 2013~2019 年数据进行建模，然后预测 2020 年数据与实测数据进行比对，总体结果如下图所示。图中上半部分给出了模型拟合/预测值与实测值过程线，下半部分给出了各分量过程线。

下面给出训练和预测结果的具体数据。

3.6.5 训练结果

使用 2013~2019 年数据进行建模，结果如下：

（1）经过预处理模块后，原因量数据如下：

处理后原因量	分量类型	测点	表达式	备注
x1	Head(水位)	Head_Up	x-75	当前上游水位-基准值（75m）的 1、2 次方
x2	Head(水位)	Head_Up	(x-75)^2	
x3	Head(水位)	Head_Up	x-75	当前下游水位-基准值（75m）的 1、2 次方
x4	Head(水位)	Head_Up	(x-75)^2	
x5	Temp(温度)	Temp_Air	x	当前气温
x6	Temp(温度)	Temp_Air	Average(x,30)	前 30 天平均气温
x7	Time(时效)	Time	ln(1+x/365.0)	

（2）原因量与效应量相关性分析结果

原因量	相关系数	偏相关系数	半偏相关系数
x1	0.811269	0.356731	0.063767
x2	0.803494	-0.21872	-0.01478
x3	-0.52788	0.188283	0.081239
x4	0.546336	0.193107	0.082617
x5	-0.82938	0.245113	0.091895
x6	-0.87787	-0.63569	-0.29197
x7	-0.21895	-0.36571	-0.13768

（3）总体建模精度

相关系数(R)	决定系数(R2)	调整后决定系数(R2-adj)	估计标准误差(stderr)
0.94	0.883	0.88	0.19558

（4）方差分析表(ANOVA)

来源	平方和	自由度	均方和	F 统计量	显著性 P
回归	67.722	7	9.675	252.922	0
残差	8.951	234	0.038		
综合	76.673	241			

（5）回归系数表(Coefficients)

原因量	非标准化系数(B)	标准误	标准化系数(Beta)	t	显著性(P>t)	相关系数	偏相关系数	半偏相关系数	偏决定系数	方差膨胀系数(VIF)
Const	-3.808	1.429		-2.665	0.008					
x1	0.139	0.03	2.519	4.648	0	0.811	0.291	0.104	1	588.787
x2	-0.001	0	-2.079	-3.82	0	0.803	-0.242	-0.085	0.999	593.559
x3	0.16	0.12	0.333	1.334	0.183	-0.528	0.087	0.03	0.984	125.188
x4	0.01	0.007	0.365	1.446	0.149	0.546	0.094	0.032	0.96	127.953
x5	0.018	0.005	0.253	3.468	0.001	-0.829	0.221	0.077	0.752	10.661
x6	-0.058	0.005	-0.76	-11.833	0	-0.878	-0.612	-0.264	0.968	8.258
x7	-0.144	0.024	-0.134	-5.955	0	-0.219	-0.363	-0.133	0.526	1.011

（6）回归方程

y = -3.80762 + 0.138915*x1 - 0.00066995*x2 + 0.159678*x3 + 0.0103074*x4 + 0.0180685*x5 - 0.058186*x6 -0.143973*x7

结合预处理后原因量信息，上述方程可以写为如下更易读的形式：

$$y=-3.80762+0.138915\left(H_{上游}-75\right)-0.00066995\left(H_{上游}-75\right)^2+0.159678\left(H_{下游}-75\right)+0.0103074\left(H_{下游}-75\right)^2$$
$$+0.0180685T-0.058186T_{30}-0.143973\ln(1+t/365)$$

（7）拟合结果

效应量实测值、模型拟合值、误差以及各分量结果见下表。

3.6.6 预测结果

利用建立的模型对 2020 年进行预测，并与实测数据比较结果如下：

（1）总体精度：决定系数=0.816921，相关系数=0.930917，标准=0.064130

（2）效应量实测值、模型拟合值、误差以及各分量结果见下表

3.7 小结

（1）统计模型是大坝安全监测中最常用和最重要的分析模型，本模型总结了位移、渗流、应力等统计模型的共性，基于通用的数据预处理模块，采用表达式形式进行因子定义，具有很强的通用性和适用性。

（2）本模型提供了多元线性回归和逐步回归两种计算方法，当原因量较多时，逐步回归可以筛选较为显著的原因量，减少回归方程的复杂度。

（3）本模型算例建模和预测结果与实测结果规律一致，各分量过程线较为合理，表明模型是有效的。模型训练结果的总体精度指标为：相关系数=0.94，决定系数=0.88，标准误差=0.196，预测结果的总体精度指标为：相关系数=0.93，决定系数=0.82，标准误差=0.06。模型精度仍有提升空间，要进一步提升建模精度，需要进一步结合工程实践经验，调整因子设置参数，并反复优化，该过程较为耗时。总之，统计模型更依赖于人工建模，在智能化方面仍有所不足。

3.8 参考文献

4. 灰色系统模型

4.1 GM(1, 1), GM(1, n)

见 12.2.1 章

4.2 DGM(1, n)

设有时间序列 $Y=\{y_i\};i=1,2,...N$

对应影响因素 $X=\{x_i^k\};i=1,2,...N,k=1,2,...h$

灰色系统的一阶，二阶，…n 阶相邻累减生成 Δ^1Y， Δ^2Y， Δ^n 定义为

$$\Delta^1Y(k)=y(k)-y(k-1)$$
$$\Delta^2Y(k)=\Delta^1Y(k)-\Delta^1Y(k-1)$$
.....
$$\Delta^nY(k)=\Delta^{n-1}Y(k)-\Delta^{n-1}Y(k-1);$$
$$k=1...N$$

当取 dgm(n, h)中 n = 1 时，对应 Δ^1Y；

灰色系统的一次相邻均值生成定义为

$$X_1=\{x_i(k)=0.5x_i(k+1)+0.5x_i(k)\};k=1...N,i=1...h$$

则可直接建立如下 n 阶 h 个变量的微分方程，记为 DGM(n ,h)模型：

$$\frac{d^nY}{dt^n}+\sum_{i=1}^{n}a_i\frac{d^{n-i}Y}{dt^{n-i}}=\sum_{i=1}^{h}b_iX_i$$

其参数列 $\hat{a}=[a_1,a_2,...,a_n,b_1,b_2,...,b_n]^T$ 用下式辨识：

$$\hat{a}=[(A\vdots B)^T(A\vdots B)]^{-1}(A\vdots B)^TZ_N$$

其中，

$$A=\begin{bmatrix}-\Delta^{n-1}Y(\frac{n}{2}+1),-\Delta^{n-2}Y(\frac{n}{2}+1),...,-\Delta Y(\frac{n}{2}+1)\\-\Delta^{n-1}Y(\frac{n}{2}+2),-\Delta^{n-2}Y(\frac{n}{2}+2),...,-\Delta Y(\frac{n}{2}+2)\\...\\-\Delta^{n-1}Y(N-\frac{n}{2}),-\Delta^{n-2}Y(N-\frac{n}{2}),...,-\Delta Y(N-\frac{n}{2})\end{bmatrix}$$

$$B=\begin{bmatrix}-X_1(\frac{n}{2}+1),-X_2(\frac{n}{2}+1),...,-X_n(\frac{n}{2}+1)\\-X_1(\frac{n}{2}+2),-X_2(\frac{n}{2}+2),...,-X_n(\frac{n}{2}+2)\\...\\-X_1(N-\frac{n}{2}),-X_2(N-\frac{n}{2}),...,-X_n(N-\frac{n}{2})\end{bmatrix}$$

常数量

$$Z_N=[\Delta^nY(\frac{n}{2}+1),\Delta^nY(\frac{n}{2}+1),...,\Delta^nY(\frac{n}{2}+1)]^T$$

证明如下:

如果原始序列等间距采样，即 $\Delta t=(t+1)-t=1$

用中心累减生成近似 Y 的导数信号 $\frac{d^nY}{dt^n}=\Delta^n(Y)$

对相应序列 x 作中心均值生成 $X_i(k)=0.5X_i(k+1)+0.5X_i(k)$

则原微分方程可变为如下等效增量模型

$$\Delta^nY+\sum_{i=1}^{n}a_i\Delta^{n-1}Y=\sum_{i=1}^{h}b_iX_i+\varepsilon$$

其中 ε 为误差项。

用最小二乘法可求得使误差项最小的参数列 $\hat{a}$ 。
证毕。

用 DGM(n, h)模型分析处理大坝检测资料时，常取 n = 1，原微分方程变为

$$\frac{d\ Y}{dt}+aY=\sum_{i=2}^{h}b_iX_i$$

由此可按照 $\hat{a}$ 的辨识式进一步求出参数列。
最终可得，离散解

$$\hat{y}_k=(y_1-\frac{1}{a}\sum_{i=1}^{h}b_ix_i^k)e^{-a(k-1)}+\frac{1}{a}\sum_{i=1}^{h}b_ix_i^k$$

5. 时间序列模型

5.1 平稳时间序列预测

5.1.1 自回归模型(AR)

为了对平稳时间序列　进行预报引入 m 阶自回归模型(AR(m)模型)

$$\hat{x}_{m+1}=\sum_{j=1}^{m}\beta_j\hat{x}_j+\varepsilon_t+c$$

式中 β_j 为自回归系数，在预报问题中，m 即为自回归模型阶数，需要指定， β_j 为待定参量；ε_t 为残差，它的均值为 0，方差为 σ_x^2 ，相关函数为 $R_x(\tau)$ ， 且当 i>0 时，有 $E(\hat{x}_{t-i}\varepsilon_\tau)=0$ 。
根据平稳时间序列的特性，可以求得相关函数的无偏或渐进无偏的统计估计值

$$r(\tau)=\frac{1}{n-\tau}\sum_{i=1}^{R-\tau}x_i\bullet x_{i+\tau},\tau=0,1,...,m$$

为求解上述参数，将　乘以预报模型公式的两边，得到方程为

$$\hat{x}_t\hat{x}_{t-i}=\sum_{j=1}^{m}\beta_j\hat{x}_{t-j}\hat{x}_{t-i}+\varepsilon_t\hat{x}_{t-i}$$

根据 $E(\hat{x}_{t-i}\varepsilon_\tau)=0$ 以及前述公式，对上式的两边取数学期望，得到自回归系数 β_j 满足 m 阶分方程

$$R(i)=\sum_{j=1}^{m}R(i-j)\beta_j$$

这里 i>0，取 i=1,2,…,m;得到 β_j 满足的 m 阶线性方程组

$$\begin{bmatrix}1 & R(1) & \cdots & R(m-1)\\ R(1) & 1 & \cdots & R(m-2)\\ \cdots & \cdots & \cdots & \cdots\\ R(m-1) & R(m-2) & \cdots & 1\end{bmatrix}\begin{Bmatrix}\beta_1\\ \beta_2\\ \cdots\\ \beta_m\end{Bmatrix}=\begin{Bmatrix}R(1)\\ R(2)\\ \cdots\\ R(m)\end{Bmatrix}$$

如果用相关函数 $R(\tau)$ 的估值 $r(\tau)$ 代替上式中的 $R(\tau)$ 得到自回归系统 β_j 的估计值 b_j ，满足 m 阶线性方程组。

用 m 阶的 Toeplitz 矩阵(r(t-j))表示时，有

$$(r(i-j))\begin{Bmatrix} b_1 \\ b_2 \\ \cdots \\ b_m \end{Bmatrix} = \begin{Bmatrix} r(1) \\ r(2) \\ \cdots \\ r(m) \end{Bmatrix}$$

上式可用 Toeplitz 矩阵(r(i-j))特有的性质地推或逐步回归算法求解。

假如在求得 m 阶线性方程组上式的解 $b_1, b_2, \cdots, b_m$ ，后，可求得(m+1)阶方程的参数

$$b_{m+1,m+1} = \frac{r(m+1) - \sum_{j=1}^{m} b_{mj} r(m+1-j)}{1 - \sum_{j=1}^{m} b_{mj}(j)}$$

$$b_{m+1,j} = b_{mj} - b_{m+1,m+1} \bullet b_{m,m+1-j}, j = 1,2,\cdots m$$

解得 b_j 后，代入 AR 模型公式，即可递归求解得到

以此类推，可得到预报值 $x_{n+1}^*, x_{n+2}^* \cdots x_{n+k}^*$ 。

5.1.2 向量自回归模型(VAR)

为了预报 K 维平稳时间序列 $\tilde{x}(t) = [x_1(t), ..., x_K(t)]^T$ ，取 m 阶自回归预报模型(VAR(m)模型)

$$\begin{bmatrix} \tilde{x}_1(t) \\ \tilde{x}_2(t) \\ \cdots \\ \tilde{x}_K(t) \end{bmatrix} = \sum_{j=1}^{m} \begin{bmatrix} g_{11}^j & g_{12}^j & \cdots & g_{1K}^j \\ g_{21}^j & g_{22}^j & \cdots & g_{2K}^j \\ \cdots & \cdots & \cdots & \cdots \\ g_{K1}^j & g_{K2}^j & \cdots & g_{KK}^j \end{bmatrix} \bullet \begin{bmatrix} \tilde{x}_1(t-j) \\ \tilde{x}_2(t-j) \\ \cdots \\ \tilde{x}_K(t-j) \end{bmatrix} + \begin{bmatrix} \varepsilon_1(t) \\ \varepsilon_2(t) \\ \cdots \\ \varepsilon_K(t) \end{bmatrix}$$

简写为

$$\tilde{x}(t) = \sum_{j=1}^{m} G_j \tilde{x}(t-j) + \varepsilon(t)$$

这里， $\tilde{x}(t)$ 是均值为 0，方差为 1 的 K 维平稳随机过程。其自相关、互相关系数 $R_{ij}(\tau) = E[\tilde{x}_i(t)\tilde{x}_j^T(t+\tau)]$ 。由 $R_{ij}(\tau)$ 定义可知， $R_{ij}(\tau) = R_{ij}(-\tau)$ 。由 $\tilde{x}(t)$ 的平稳特性可以得到 $R_{ij}(\tau)$ 的无偏或渐进无偏统计估计量 $r_{ij}(\tau) = \frac{1}{n-\tau} \sum_{k=1}^{n-\tau} x_i(k) x_j(k+\tau)$

对应的， $\tilde{x}(t)$ 相关系数矩阵为

$$R(t) = E[\tilde{x}(t)\tilde{x}^T(t+\tau)] = \begin{bmatrix} R_{11}(\tau) & R_{12}(\tau) & \cdots & R_{1K}(\tau) \\ R_{21}(\tau) & R_{22}(\tau) & \cdots & R_{2K}(\tau) \\ \cdots & \cdots & \cdots & \cdots \\ R_{K1}(\tau) & R_{K2}(\tau) & \cdots & R_{KK}(\tau) \end{bmatrix}$$

可知，$R^T(\tau)=R(-\tau)$

残差 $\varepsilon(t)=[\varepsilon_1(t),...,\varepsilon_K(t)]^T$。特性可知，其相关函数

其中回归系数矩阵

$$G_j=\begin{bmatrix} g_{11}^j & g_{12}^j & \cdots & g_{1K}^j \\ g_{21}^j & g_{22}^j & \cdots & g_{2K}^j \\ \cdots & \cdots & \cdots & \cdots \\ g_{K1}^j & g_{K2}^j & \cdots & g_{KK}^j \end{bmatrix}$$

是进行预报的待定参量

用 $\tilde{x}^T(t-i)$ 乘以上述公式两边，并取数学期望，得到回归系数矩阵 G_j 满足 m 阶差分方程

$$R(-i)=\sum_{j=1}^{m}G_jR(j-i)$$

将上式的两端转置，得到回归系统矩阵 G_j 满足 K 个 m 阶差分方程

$$R(i)=\sum_{j=1}^{m}R(j-i)G_j^T,i=1,2,...,K$$

把计算得到的关系数 $R_{ij}(\tau)$ 的估计量 $r_{ij}(\tau)$ 代入上式，求解回归系数矩阵 G_j 的估计矩阵，可得到 $\tilde{x}(n+1)=\sum_{j=1}^{m}B_j\tilde{x}(n+1-j)$ 的预报值

5.2 非平稳时间序列分析

在实际问题中遇到的多数物理数据一般都是非平稳的，AR 模型只能处理平稳数据，因此，对非平稳的时间序列进行分析面具有更大的实际意义。对非平稳时间序列进行分析时，可以使用参数模型法，将其分解为主值函数项（不平稳）和周期函数项（平稳）。分解方法一般有加法模型和乘法模型。

5.2.1 加法模型

为预报非平稳时间序列 $x(t)$ ，加法模型将 $x(t)$ 分解为 $\varphi(t)$ 与 $\eta(t)$ 之和：

$$x(t)=\varphi(t)+\eta(t)=f(t)+p(t)+\eta(t)$$

其中 $\varphi(t)$ 为趋势函数项，包含主值函数项 $f(t)$ 和周期函数项 $p(t)$；$\eta(t)$ 为剩余部分。

由 $x(t)$ 的测量数据 $x_1,x_2,\cdots,x_j,\cdots x_n$，用统计分析方法识别、提取趋势函数 $\varphi(t)$ ，并估计其参数，是参数模型的基本思路。

下面接收计算各项的方法

(1) 计算主值函数项 $f(t)$

当不考虑影响因素而只考虑主值函数随时间的变化时，可以取 $f(t)$ 的形式为

$$f(t)=a_0+\sum_{i=1}^{4}a_it^i+a_5t^{-1}+a_6t^{-2}+a_7t^{1/2}+a_8t^{-1/2}+a_9e^{-t}+a_{10}\ln t$$

当考虑每个影响因素时，若有 K 个影响因素，即 $z(t)=\begin{bmatrix} z_t^1 & z_t^2 & \cdots & z_t^K \end{bmatrix}$，可以宽泛地取 $f(t)$ 形式为

$$f(t)=a_0+\sum_{i=1}^{K}a_i z_t^{\mathrm{i}}$$

特别的，在该情况下，可以将水深、水深的二次方、三次方、四次方，温度，时间及时间的对数等作为影响因素传入，具有更好的统计学意义。

根据观测值 $x(t)$，使用逐步回归法挑选 $f(t)$ 的参数 a_i。若 $a_i=0(i=1,2,...)$，说明该时间序列无影响因素。

(2) 计算周期函数项 $p(t)$

去掉 $f(t)$ 后，可以得到一个新的时间序列：$q(t)=x(t)-f(t)$

对这组数据考虑一个隐含周期模型

$$q(t)=P(t)+\eta(t)$$

$$P(t)=b_0+\sum_{j=1}^{t}(b_{1j}\cos(w_j t)+b_{2j}\sin(w_j t)$$

其角速度 $w_j=\dfrac{2\pi}{T_j}$

下面通过统计方法得到可能的周期 T_j

首先对于所有可能周期 $T_j=n/k,k=1,2,...K$ ，计算其振幅

$$b_{1k}=\frac{2}{n}\sum_{j=1}^{k}q_j\cos\frac{2\pi}{n}k_j$$

$$b_{2k=}\frac{2}{n}\sum_{j=1}^{k}q_j\sin\frac{2\pi}{n}k_j$$

其中 $k=1,2,...\left\lceil\dfrac{n-1}{2}\right\rceil$

取统计量 $S_k^2=\dfrac{1}{2}(b_{1k}^2+b_{2k}^2),k=1,2,...,\left\lceil\dfrac{n-1}{2}\right\rceil$

继续取 $S^2=\sum_{k=1}^{K}S_k^2$

为了从这些周期中选取出随机过程 $x(t)$ 真正周期，设 S_i^2 为集合 $\{S_1^2,S_2^2,...,S_K^2\}$ 中的第 i 个最大值

在 $\eta(t)$ 为高斯白噪声的假定系，统计量 $y_i=S_i^2/S^2$ 服从 Fisher 分布

$$P\{y>y_i\}=\sum_{j=0}^{r}(-1)^j C_k^{j+1}[1-(j+1)y_i]^{k-1}$$

其中 r 是使 $1-(r+1)y_i>0$ 成立的最大整数

对给定的显著水平 a，若 $P\{y>y_i\}\geq a$，则认为随机过程 $x(t)$ 无周期函数项 P(t)。否则，以显著水平 a，接受 S_i^2 对应的周期 T_i 为随机过程的第 i 个周期。

得到所有可能的周期后，求出对应的 w_j，代入 $p(t)$ 表达式中，即可得到周期项 $p(t)$。

(3) 计算剩余项 $\eta(t)$

在求得 $\varphi(t)$ 后，得到时间序列 $x(t)$ 的趋势项函数。那么，从 $x(t)$ 中提取趋势函数项后，剩余部分为 $\eta(t)=x(t)-\varphi(t)$

由于去掉趋势函数项，因此 $\eta(t)$ 可作为一个平稳随机过程，按照前面介绍的平稳随机过程处理方法进行分析和预报

5.2.2 乘法模型

与加法模型不同，为预报非平稳时间序列 $x(t)$ ，乘法模型将 $x(t)$ 分解为 $\varphi(t)$ 与 $\eta(t)$ 之积：

$$x(t)=\varphi(t)\bullet\eta(t)=f(t)\bullet p(t)\bullet\eta(t)$$

其中各项含义与加法模型相同。

使用乘法模型求解时，可按照加法模型相同的方法先解得 $f(t)$，则 $q(t)=x(t)/f(t)$，且有 $q(t)=P(t)\bullet\eta(t)$。按照加法模型相同方法求解 $P(t)$ 后，可得到 $\varphi(t)$，进而得到

$$\eta(t)=x(t)/\varphi(t)$$

此时 $\eta(t)$ 为平稳时间序列，按照前面的方法继续求解即可。

6. 卡尔曼滤波模型

7. 神经网络模型与遗传算法神经网络模型

8. 支持向量机模型与遗传算法支持向量机模型

8.1 支持向量机模型

见 12.2.3

8.2 遗传算法优化的支持向量机

遗传算法优化支持向量机

使用遗传算法优化支持向量机也就是优化支持向量机的各种可调参数。遗传算法优化过程中，种群中每个个体代表了问题的一个可能解。遗传算法单次操作过程分为以下几个步骤：选择；交叉；变异。操作过程中，将每个个体的所有参数看作其对应的染色体，每个参数为染色体的一段。遗传算法优化支持向量机的流程如下：

(1) 初始化

设置需要使用遗传算法优化的所有参数及其取值范围，设置种群内个体数目 pop_size,种群最大代数 max_pop,设置交叉权重 cross_weight,交叉概率 cross_prob,设置变异概率 mutation_prob。按照这些参数构建 pop_size 个 SVR 模型，作为初代种群。并设置训练数据和验证数据。设置交叉权重。

(2) 训练种群

对种群内的每个 SVR 模型，使用训练数据进行训练。

(3) 选择操作

对每个 SVR 个体，用模型对验证数据进行预测，并使用预测结果与样本真实值的 R2_Score 作为其适应度：

$$R2\_Score = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

其中 y_i 为第 i 个样本真实值，$\hat{y}_i$ 为模型对第 i 个样本预测值，$\bar{y}$ 为样本均值。R2_Score 越高，表示模型对样本拟合的越好。R2_Score=1,表示预测值和真实值完全相同

计算出所有个体适应度后，采用轮盘赌算法，选择出 pop_size 数目的个体，作为下一代种群，该过程中适应度高的个体有更大几率被选中。由于 R2_Score 可能为负，此时需要对种群所有个体的适应度加上一个值，使适应度均为正值。

(4) 交叉操作

随机选择两个个体，设其染色体为 A，B，按照个体间交叉概率 cross_prob 对 A，B 进行交叉操作：

随机选择交叉点，随机选择向前或向后的交叉方向（当向后交叉时，对两个个体的染色体交叉点以后的部分操作；向前交叉时，对两个个体的染色体在交叉点以前的部分进行操作）

对交叉方向上 A，B 的每一段 va, vb，按照交叉权重重新设置值：

$$va = va * cross\_weight + (1 - cross\_weight) * vb$$
$$vb = vb * cross\_weight + (1 - cross\_weight) * va$$

其中 cross_weight 为指定的交叉权重

(5) 变异操作

按照指定的变异概率 mutation_prob 对种群内每个个体的染色体 A 进行变异操作：

随机选择 A 上的一段 va，再随机选择两个个 0 到 1 之间的值 p，q，计算新的 va 值:

$$va = va - \left(va - \min\_va\right) * \left(1 - q^{1 - \frac{t}{\max\_pop}}\right)$$
$$va = va - \left(\max\_va - va\right) * \left(1 - q^{1 - \frac{t}{\max\_pop}}\right)$$

其中 min_va, max_va 分别表示 va 的最小值和最大值，t 和 max_pop 表示当前种群代数和种群最大代数。该变异公式使得计算轮次越大，变异趋向于更小的幅度。

先进行上面的第（1）步，然后重复第（2）-（5）max_pop 次，选择出种群中适应度最高的个体（SVR 模型），作为最优解。

9. 极限学习机模型

极限学习机在结构上，就是一个向前传播的神经网络，但其输入层和隐含层的连接权值、隐含层的阈值可以随机设定，且设定完后不用再调整。而 BP 神经网络需要不断通过反向传播去调整权值和阈值。同时，隐含层和输出层之间的连接权值不需要迭代调整，而是通过解方程组一次性确定，减少了计算量。其基本结构如下图所示。

假设有 N 个任意的样本(X_i, t_i)，其中$\mathrm{X}_i = [x_{i1}, x_{i2}, ..., x_{in}]^T \in R^n, t_i = [t_{i1}, t_{i2}, ...t_{im}]^T \in R^m$

对于一个有 L 个隐层节点的单隐层神经网络可以表示为

$$\sum_{i=1}^{L} \beta_i g(W_i \cdot X_j + b_i) = o_j, j = 1, 2, ...N$$

其中$W_i = [w_{i1}, w_{i2,}...w_{in}]^T$为输入权重，$\beta_i$为输出权重，$b_i$是第 i 个隐层单元的偏置。g 表示激活函数（也称为特征映射），常用激活函数有三角函数，高斯函数，径向基函数，Sigmoid 函数等。

单隐层神经网络学习的目标是使得输出的误差最小，可以表示为

$$\sum_{j=1}^{N} \| o_j - t_j \| = 0$$

即存在β_i，W_i和b_i，使得

$$\sum_{i=1}^{L} \beta_i g(W_i \cdot X_j + b_i) = t_j, j = 1, 2, ...N$$

可以使用矩阵表示为

$$H\beta = T$$

其中，H是隐节点的输出，β为输出权重，T为期望输出

即：

$$H(W_1, W_2, ...W_L, b_1, b_2, ..., b_L, X_1, X_2, ...X_L)$$

$$= \begin{bmatrix} g(W_1 \cdot X_1 + b_1) \cdots g(W_L \cdot X_1 + b_L) \\ \vdots \quad \cdots \quad \vdots \\ g(W_1 \cdot X_1 + b_1) \cdots g(W_L \cdot X_1 + b_L) \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_2^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L \times m}$$

$$T = \begin{bmatrix} T_1^T \\ T_2^T \\ \vdots \\ T_N^T \end{bmatrix}_{N \times m}$$

为了能够训练单隐层神经网络，需要计算$\hat{W}_i$，$\hat{b}_i$和$\hat{\beta}_i$，使得

$$\left\| H(\hat{W}_i, \hat{b}_i)\hat{\beta}_i - T \right\| = \min_{W,b,\beta} \left\| H(W_i, b_i)\beta_i - T \right\|, i = 1, ...L$$

这等价于

$$E = \sum_{j=1}^{N} (\sum_{i=1}^{L} \beta_i g(W_i \cdot X_j + b_i) - t_j)^2$$

在 ELM 算法中，一旦输入权重W_i和隐层的偏置β_i被随机确定，隐层的输出矩阵H就被唯一确定。

训练单隐层神经网络可以转化为求解一个线性系统 $H\beta = T$ 。且其最小二乘解为 $\hat{\beta}_i = H^+T$

其中， H^+ 为矩阵 H 的 Moore-Penrose 广义逆。

10. 模糊预测模型

目前使用较多的模糊聚类方法为模糊 C 均值聚类算法（FCM）。

若对于论域 D 中任一元素 x，都有一个数 $U(x) \in [0,1]$ 与之对应，则称 U 为 D 上的模糊集， $U(x)$ 称为 x 对 D 的隶属度。 $U(x)$ 越接近于 1，表示 x 属于 U 的程度越高， $U(x)$ 越接近于 0 表示 x 属于 U 的程度越低。

FCM 使用一个隶属度矩阵 $U = [u_{is}]_{N\times K}$ 表示每个数据属于每个类的程度大小，其中 N 是数据集大小，K 是聚类个数（聚类中心个数），且有 $\sum_{i=1}^{K} u_{is} = 1$ 。

模糊聚类的目标函数：

$$J(U,C) = \sum_{s=1}^{N}\sum_{i=1}^{K}(u_{is})^m * dist(c_i, x_s)^2$$

其中,C 为聚类中心，m 为加权指数， $dist(c_i, x_s)^2$ 即 c_i, x_s 欧氏距离的平方。也就是说，模糊聚类的目标函数 $J(U,C)$ 就是各个数据点到每个聚类中心的加权平方和。

模糊聚类除了计算聚类中心 C 之外，并不会将数据点直接归类，而是计算隶属度矩阵 U，取每条数据中，可能性最大的聚类中心作为对应的最终类别。

上述问题即为：在隶属度 $\sum_{i=1}^{K} u_{is} = 1$ 的约束条件下，最小化 $J(U,C)$ ，即：

$$\min\{J(U,C)\} = \min\{\sum_{s=1}^{N}\sum_{i=1}^{K}(u_{is})^m * dist(c_i, x_s)^2\}$$

使用拉格朗日方法可以解得：

$$u_{is} = \frac{1}{\sum_{j=1}^{K}(\frac{dist_{is}^2}{dist_{js}^2})^{-\frac{2}{m-1}}}$$

进一步代入并求解，得到

$$c_i = \frac{\sum_{s=1}^{N} u_{is}^m x_s}{\sum_{s=1}^{N} u_{is}^m}$$

可知 U 和聚类中心 C 相互关联，于是在算法开始时先给其中一个变量赋值，再计算得到另一个变量，由此不断迭代和更新，使目标函数 $J(U,C)$ 不断减小，最终达到一个稳定条件，即，最终两次迭代的 U 小于误差值 ε ：

$$|U^{k+1} - U^k| < \varepsilon$$

在使用模糊聚类算法进行大坝位移预测时，按如下步骤进行：

1. 需要先指定聚类中心数目 K。对于输入数据 U，先得到因变量的数据范围 $r=\max(U_y)-\min(U_y)$，即因变量的最大值与最小值之差。
2. 按照 $\frac{r}{K}$ 的间隔，依次划分出 K 个区间，再按照因变量的范围，将数据标注为具体的类别
3. 将标注好的输入 FCM 模型中训练
4. 预测时，输入自变量，对每条数据，根据模型输出的类别，得到因变量对应的模糊区间

11. 小波分析功能

12. GM-BP-SVM 智能组合模型

12.1 功能概述

由于实际大坝工作条件复杂，影响因素众多，还有许多未知或者不确定因素，因此原因量和效应量之间存在非常复杂的关系，各种分析模型往往有擅长的情况，也有不理想的情况，很难有单一模型能够适用于所有问题。针对这一问题，本组合模型采用最优加权组合法，将灰色系统模型（GM）、神经网络模型（BP）和支持向量机模型（SVM）有机结合，取长补短，以获得更优的建模和预测效果。

12.2 基本理论

12.2.1 灰色系统基本理论

客观世界是物质的世界，也是信息的世界。既有大量已知信息，也有不少未知的或非确知的信息。未知或非确知信息称之为黑色的；已知信息称之为白色的。既含有已知信息又含有未知或不确定信息的系统称为灰色系统。从另一方面讲，对一个客体（例如承受水压力、变温等荷载作用的大坝）可以用一组状态变量构成的系统来描述，此系统可视为“箱”。黑箱指的是输入和输出关系而不计内部结构与状态的系统；或指内部状态不可监测、不可控制的系统。相反，对内部结构完全可以监测和控制的系统称为白箱。而介于这两者之间的就是灰箱或灰色系统。研究灰色系统的理论就是灰色系统理论，简称 GS(Grey System)理论。

坝体和基础组成的大坝是一个灰色系统，通过监测得到的较少信息，建立所需微分方程的动态模型。在此基础上进行分析，进一步认识大坝原型结构特性及其稳定性。将大坝监测效应量当做一定范围内变化的灰色量，将其监测的资料视为一定时区变化的灰色过程，将无或弱规律变化的原始数列变为有较强规律变化的生成数据，并以此建立灰色模型，这种模型实际上是生成数据模型，简称 GM(Grey Model)模型。然后将模型的计算值进行逆生还原为原始数据进行预测效应量的变化规律和演变趋势。根据以上思路，用灰色理论建模的基本原理概述如下。

设给定时间序列：

$$\left\{x_i^{(0)}\left(t_j\right)\right\};i=1,2,...,M\left(\text{因子数}\right);\ j=1,2,...,N\left(\text{样本量}\right) \tag{12.1}$$

有相应的一阶累加序列：

$$\left\{x_i^{(1)}\left(t_j\right)\right\};i=1,2,...,M;\ j=1,2,...,N;\ x_i^{(1)}=\sum_{s=1}^{j}x_i^{(0)}\left(t_s\right) \tag{12.2}$$

有相应的多次累差序列：

$$\left\{a^{(k)}(x_i,j)\right\}\quad i=1,2,...,M;\ \ j=2,3,...,N;\ \ k=1,2,...,n$$

$$\begin{aligned}
&a^{(1)}(x_i,j)=x_i^{(0)}(t_j)\\
&a^{(2)}(x_i,j)=x_i^{(0)}(t_j)-x_i^{(0)}(t_{j-1})\\
&.....\\
&a^{(k)}(x_i,j)=a^{(k-1)}(x_i,j)-a^{(k-1)}(x_i,j-1)\ \ (k=3,4,...,n)
\end{aligned}\tag{12.3}$$

则作如下数据处理，并采用等价记法 $x_i(t_j)=x_i(j)$

$$A=\begin{pmatrix}
-a^{(n-1)}\left(x_1^{(1)},2\right) & -a^{(n-2)}\left(x_1^{(1)},2\right) & \cdots & -a^{(1)}\left(x_1^{(1)},2\right)\\
-a^{(n-1)}\left(x_1^{(1)},3\right) & -a^{(n-2)}\left(x_1^{(1)},3\right) & \cdots & -a^{(1)}\left(x_1^{(1)},3\right)\\
\vdots & \vdots & & \vdots\\
-a^{(n-1)}\left(x_1^{(1)},N\right) & -a^{(n-2)}\left(x_1^{(1)},N\right) & \cdots & -a^{(1)}\left(x_1^{(1)},N\right)
\end{pmatrix}$$

$$B=\begin{pmatrix}
-\frac{1}{2}\left(x_1^{(1)}(2)+x_1^{(1)}(1)\right) & x_2^{(1)}(2) & \cdots & x_M^{(1)}(2)\\
-\frac{1}{2}\left(x_1^{(1)}(3)+x_1^{(1)}(2)\right) & x_2^{(1)}(3) & \cdots & x_M^{(1)}(3)\\
\vdots & \vdots & & \vdots\\
-\frac{1}{2}\left(x_1^{(1)}(N)+x_1^{(1)}(N-1)\right) & x_2^{(1)}(N) & \cdots & x_M^{(1)}(N)
\end{pmatrix}$$

$$\hat{a}=\begin{bmatrix}a_1 & a_2 & \cdots & a_n & : & b_1 & \cdots & b_{M-1}\end{bmatrix}^T$$

$$Y_N=\begin{bmatrix}a^{(n)}\left(x_1^{(1)},2\right) & a^{(n)}\left(x_1^{(1)},3\right) & \cdots & a^{(n)}\left(x_1^{(1)},N\right)\end{bmatrix}^T$$

在定义灰色导数的基础上，建立分析资料序列式(12.2)或式(12.3)的变化趋势的微分方程为：

$$\sum_{i=0}^{n}a_i\frac{d^{(n-i)}\left(x_1^{(1)}\right)}{dt^{n-i}}=\sum_{i=1}^{M-1}b_ix_{i+1}^{(1)}\tag{12.4}$$

将上式化为差分方程后：

$$\sum_{i=0}^{n}a_i\frac{\Delta^{(n-i)}\left(x_1^{(1)}(t)\right)}{\Delta t^{n-i}}=\sum_{i=1}^{M-1}b_ix_{i+1}^{(1)}\tag{12.4}$$

对具有物理、力学等意义的灰色系统，为非负时间序列。可令 $\Delta t=1$，且 $a_0=1$，从而有：

$$\sum_{i=0}^{n}a_ia^{(n-i)}\left(x_1^{(1)},j\right)=\sum_{i=1}^{M-1}b_ix_{i+1}^{(1)}(t_j)\ \ (j=2,...,N)$$

即：

$$\begin{bmatrix}a^{(n)}\left(x_1^{(1)},2\right)\\ \vdots\\ a^{(n)}\left(x_1^{(1)},N\right)\end{bmatrix}=-\begin{pmatrix}
-a^{(n-1)}\left(x_1^{(1)},2\right) & \cdots & a^{(1)}\left(x_1^{(1)},2\right)\\
\vdots & & \vdots\\
-a^{(n-1)}\left(x_1^{(1)},N\right) & \cdots & a^{(1)}\left(x_1^{(1)},N\right)
\end{pmatrix}$$

$$\begin{pmatrix}a_1\\ \vdots\\ a_{n-1}\end{pmatrix}+\begin{pmatrix}
-a^{(0)}\left(x_1^{(1)},2\right) & x_2^{(1)}(2) & \cdots & x_M^{(1)}(2)\\
\vdots & \vdots & & \vdots\\
-a^{(0)}\left(x_1^{(1)},N\right) & x_2^{(1)}(N) & \cdots & x_M^{(1)}(N)
\end{pmatrix}\tag{12.5}$$

应用灰数生成理论按最小二乘法对 a 求解，可得：

$$\hat{a}=\left(\left[A:B\right]^{T}\left[A:B\right]\right)^{-1}\left[A:B\right]^{T}Y_{N} \tag{12.6}$$

阶数 n 选择不同，建立微分方程的阶数也就不同，一般可取 n=1 时的模型 GM(1,1)和 GM(1,N)

12.2.2 BP 神经网络基本理论

人工神经网络（Artificial Neural Networks ，简写为 ANNs ）也简称为神经网络（ ANNs ）或称作连接模型 Connection Model ），它是一种模范动物神经网络行为特征，进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间相互连接的关系，从而达到处理信息的目的。神经元网络是机器学习学科中的一个重要部分，用来 classification 或者 regression 。思维学普遍认为，人类大脑的思维分为抽象（逻辑）思维、形象（直观）思维和灵感（顿悟）思维三种基本方式。逻辑性的思维是指根据逻辑规则进行推理的过程；它先将信息化成概念，并用符号表示，然后，根据符号运算按串行模式进行逻辑推理；这一过程可以写成串行的指令，让计算机执行。然而，直观性的思维是将分布式存储的信息综合起来，结果是忽然间产生想法或解决问题的办法。这种思维方式的根本之点在于以下两点：1）信息是通过神经元上的兴奋模式分布存储在网络上；2）信息处理是通过神经元之间同时相互作用的动态过程来完成的。

人工神经网络就是模拟人思维的第二种方式。这是一个非线性动力学系统，其特色在于信息的分布式存储和并行协同处理。虽然单个神经元的结构极其简单，功能有限，但大量神经元构成的网络系统所能实现的行为却是极其丰富多彩的。

（1）BP 网络模型处理信息的基本原理

输入信号 X_i 通过中间节点（隐层点）作用于输出节点，经过非线性变换，产生输出信号 Y_k，网络训练的每个样本包括输入向量 X 和期望输出量 t，网络输出值 Y 与期望输出值 t 之间的偏差，通过调整输入节点与隐层节点的联接强度取值 W_{ij} 和隐层节点与输出节点之间的联接强度 T_{jk} 以及阈值，使误差沿梯度方向下降，经过反复学习训练，确定与最小误差相对应的网络参数（权值和阈值），训练即告停止。此时经过训练的神经网络即能对类似样本的输入信息，自行处理输出误差最小的经过非线性转换的信息。

（2）BP 神经网络的确定

输入层、隐含层和输出层，隐含层的节点数小于输入节点数，输入节点数与输出节点数相同。在工作过程中，输入信号通过隐含层点作用于输出点，经过非线性变换产生输出信号，通过调整输入节点和隐含层节点的联接强度，使输出模式尽可能的等于输入模式。输入模式将网络数据通过少量的隐含层单元映射到输出模式。当隐含层的单元数比输入模式少时，就意味着隐含层就能更有效的表现输入模式，并把这种表现传送给输出层，输出层节点数和输出层节点数相同。单个隐含层的网络可以通过适当增加神经元节点的个数实现任意非线性映射，所以点个隐含层可以满足大部分的应用。当隐含层神经元的个数较少时，就意味着隐含层能用更少的数来表现输入模式。因此在三层 BP 神经网络技术中，要通过控制隐含层的数量，来达到使输出层的个数与输入层的个数相同。

（3）BP 神经网络模型创建

建立由输入层、输出层和隐含层 构成的多层感知器的 BP 神经网络，确定网络的输入层节点数 n、隐含层节点数 l，隐含层激励函数 f，输出层节点数 m，将输入层与隐含层之间的连接权值 w_{ij} 及隐含层阈值 a_j 初始化，确定学习速率和激活函数等参数。将样本数据送入输入节点，经隐含层逐层

处理后，由下式计算隐含层输出 H_j。

$$H_j = f\left(\sum_{i=1}^{n} w_{ij}x_i - a_j\right) \quad j=1,2,...,l$$

将隐含层与输出层之间的权值 w_{jk} 及输出层阈值 b_k 初始化，计算 BP 神经网络预测输出 O_k。

$$O_k = \sum_{j=1}^{l} H_j w_{jk} - b_k \quad k=1,2,...,m$$

根据网络预测输出 O_k 和期望输出 Y_k，按下式计算网络预测误差。计算每次输入样本数据的误差值，若两者误差相对较大，则误差开始反向传输，并调整神经网络参数，直到系统收敛。

$$e_k = Y_k - O_k \quad k=1,2,...,m$$

（4）BP 神经网络预测：通过网络训练，根据网络预测误差 e 更新权值 w_{ij}，w_{jk}

$$w_{ij} = w_{ij} + \eta H_j\left(1-H_j\right)x(i)\sum_{k=1}^{m} w_{jk}e_k \quad i=1,2,...,n;\ j=1,2,...,l$$

$$w_{jk} = w_{jk} + \eta H_j e_k \quad j=1,2,...,l;\ k=1,2,...,m$$

更新网络节点阈值 a_j 和 b_k

$$a_j = a_j + \eta H_j\left(1-H_j\right)x(x)\sum_{k=1}^{m} w_{jk}e_k \quad j=1,2,...,l$$

$$b_k = b_k + e_k \quad k=1,2,...,m$$

权值确定后，将数据输入到模型中进行结果预测。

12.2.3 SVM 支持向量机基本理论

支持向量机多用于分类问题，但也可以用于回归，此时通常称为支持向量回归模型（SVR）。该模型最终的模型函数表示为：

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$$

其中 $\mathbf{w}$，b 是模型要求解的主要系数。

对于给定的训练样本 $D=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),...,(\mathbf{x}_m,y_m)\}$ 传统回归模型往往直接基于模型输出和真实值之间的差别来计算损失，而 SVR 仅当 $f(\mathbf{x})$ 与 y 之间的差别大于 ε 时才计算损失，参数 ε 由用户指定。于是，根据支持向量机模型的其它理论，SVR 问题的优化目标写为：

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m} l(f(\mathbf{x}_i) - y_i)$$

其中 C 是损失系数，可由用户指定或通过遗传算法优化得到；而 l 是 ε 不敏感损失函数，仅当 f(x) 与 y 之间的差别大于 ε 时才计算损失：

$$l(z) = \begin{cases} 0, & |z| \le \varepsilon \\ |z| - \varepsilon, & |z| > \varepsilon \end{cases}$$

引入松弛变量 $\xi, \hat{\xi}_i$，则上式可重写为

$$\min_{\mathbf{w},b,\xi_i,\hat{\xi}_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\left(\xi_i + \hat{\xi}_i\right)$$

其中

$$\begin{cases}\xi_i = y_i - \left(f(\mathbf{x}_i)+\varepsilon\right), & y_i > f(x_i)+\varepsilon \\ \xi_i = 0, & y_i \le f(x_i)-\varepsilon\end{cases}$$

$$\begin{cases}\hat{\xi}_i = \left(f(\mathbf{x}_i)-\varepsilon\right)-y_i & y_i < f(x_i)-\varepsilon \\ \hat{\xi}_i = 0, & y_i \ge f(x_i)-\varepsilon\end{cases}$$

针对上述问题，首先引入拉格朗日乘子：

$$u_i \ge 0, u_i^* \ge 0, \alpha_i \ge 0, \alpha_i^* \ge 0$$

构建拉格朗日函数：

$$L\left(\mathbf{w},b,\xi,\hat{\xi},\alpha,\alpha^*,u,u^*\right) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\left(\xi_i+\hat{\xi}_i\right) + \sum_{i=1}^{m}\alpha_i\left(f(\mathbf{x}_i)-y_i-\varepsilon-\xi_i\right) + \sum_{i=1}^{m}\alpha_i^*\left(y_i-f(\mathbf{x}_i)-\varepsilon-\hat{\xi}_i\right) + \sum_{i=1}^{m}u_i(0-\xi_i) + \sum_{i=1}^{m}u_i^*\left(0-\hat{\xi}_i\right)$$

为求该函数对于$\mathbf{w},b,\xi,\hat{\xi}$的极小值，分别对$\mathbf{w},b,\xi,\hat{\xi}$求偏导，并令偏导为 0，可得：

$$\mathbf{w} = \sum_{i=1}^{m}\left(\alpha_i^*-\alpha_i\right)\mathbf{x}_i$$

$$0 = \sum_{i=1}^{m}\left(\alpha_i^*-\alpha_i\right)$$

$$C = \alpha_i + u_i$$

$$C = \alpha_i^* + u_i^*$$

则 SVR 对应的偶问题是：

$$\max_{\alpha,\alpha^*,u,u^*} L\left(\mathbf{w},b,\xi,\hat{\xi},\alpha,\alpha^*,u,u^*\right)$$

该对偶问题有解得充要条件是满足 KKT 条件，其 KKT 条件如下：

$$\begin{cases}\alpha_i\left(f(\mathbf{x}_i)-y_i-\varepsilon-\xi_i\right)=0 \\ \alpha_i^*\left(y_i-f(\mathbf{x}_i)-\varepsilon-\hat{\xi}_i\right)=0 \\ \alpha_i\alpha_i^*=0 \\ \xi_i*\hat{\xi}_i=0 \\ \left(C-\alpha_i\right)\xi_i=0 \\ \left(C-\alpha_i^*\right)\hat{\xi}_i=0\end{cases}$$

将求偏导后所得各式带入对偶式，并进行化简，可得化简后的对偶式：

$$\max_{\alpha,\alpha^*}\sum_{i=1}^{m}y_i\left(\alpha_i^*-\alpha_i\right)-\varepsilon\left(\alpha_i^*+\alpha_i\right)-\frac{1}{2}\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\alpha_i^*-\alpha_i\right)\left(\alpha_i^*-\alpha_j\right)\mathbf{x}^T{}_i\mathbf{x}_j\right)$$

然后可 SMO 算法求解该问题，求解之前，先要将α_i^*与α_i化成一个系数，因为 SMO 算法针对的是样本x_i仅有一个参数α_i的情况。对此，从 KKT 条件的前两个式子中可以推断，α_i和 α_i^*至少有一个为 0，不妨设$\lambda_i = \alpha_i - \alpha_i^*$，则$|\lambda_i| = \alpha_i + \alpha_i^*$， 则原式继续化简为只有一个变量$\lambda$的形式，从而可以使用 SMO 算法求解$\lambda$，并进一步可以得到$\alpha-\alpha^*$的值，进而得到 w 的值。

由 KKT 条件第五式可以看出，若$0<\alpha_i<C$，且由$\alpha_i(f(\mathbf{x}_i)-y_i-\varepsilon-\xi_i)=0$，则必有$\xi_i=0$，进

而有：

$$b = y_i + \varepsilon - \mathbf{w}^T \mathbf{x}_i$$

在具体实现时可计算所有满足$0 < \alpha_i < C$条件的样本求解 b 后取平均值。

12.2.4 最优加权组合理论

设构造了m个单一的监测预报模型k_i（i=1，2，…m），构成组合预报模型K_q=φ（k_1，k_2，…k_n），其中$n \le m$，q为模型的组合数。

设组合模型中各单一模型的权向量P=[p_1，p_2，…p_n]，并取$\sum_{j=1}^{n} p_j = 1$，此时组合预报模型的形式为：

$$K = p_1\hat{K}_1 + p_2\hat{K}_2 + \cdots + p_n\hat{K}_n = \sum_{j=1}^{n} p_j\hat{K}_j$$

设某个单一模型的拟合残差为：

$$e_{ti} = k_{ti} - \hat{K}_{tj} \quad (j = 1,2,...m;\ t = 1,2,...,n)$$

则各单一预报模型可构成拟合残差矩阵：

$$V = \left[\sum_{t=1}^{n} e_{ti}e_{tj}\right] \quad (i, j = 1,2,...m)$$

按最小二乘原理求解目标函数：

$$\begin{cases} Q = \sum_{i=1}^{n} e_i^2 = \min \\ s.t = \sum_{j=1}^{n} p_j = 1 \end{cases}$$

令R=[1，1，…，1]^T，则上式成为：

$$\begin{cases} Q = \sum_{i=1}^{n} e_i^2 = P^T V P = \min \\ s.t = \sum_{j=1}^{n} p_j = R^T P = 1 \end{cases}$$

求解得最优权重向量为：

$$P_0 = \frac{V^{-1}R}{R^T V^{-1} R}$$

由上式即可解得组合预测模型中各个单一模型的最优权重比。

组合预报模型比任何单一预报模型有较低的均方差，在最大化信息利用的基础上集合了单一监测模型包含的所有信息，在大多数情况下，将单一监测预报模型通过科学的组合，可以提高大坝变形监控预报的精度。

12.3 模型接口

本模型提供训练和预测两个接口，为 WebAPI 形式，通过 POST 发送 JSON 格式请求到对应地址进行调用，具体地址如下：

训练：/AnalysisModel/GM_BM_SVM/Train

预测：/AnalysisModel/GM_BM_SVM/Predict

接口输入数据格式由数据预处理模块定义，详见 2.2 节。原因量（xData，xCol）、效应量（yData，yCol）、因子设置（Factor）和模型选项（Setting）均采用统一的模式。调用模型训练接口时，需要提

供原因量、效应量、因子设置和模型选项四个方面的数据和参数，调用模型预测接口时，有如下要点：

（1）预测用的原因量必须跟训练用的原因量完全对应，例如训练时采用了气温作为原因量，那么预测时必须提供气温数据，否则程序无法执行；如果预测时额外提供了训练时没有用到的数据，则程序会自行忽略；

（2）模型训练时，程序会保存相应的因子设置和模型选项，因此预测时不需要提供因子设置和模型选项，程序会根据保存的信息自动执行因子处理，从原始原因量获得处理后原因量，该过程与训练时的因子处理过程一致。

另外需要注意，由于 GM 模型要求数据时间间隔相等，因此建议输入数据满足等时间间隔的条件，如不满足程序会计算平均时间间隔，然后通过线性插值获得等时间间隔的数据，确保模型能够顺利运行。另外在预测时，GM 模型要求预测起始时间必须紧接训练结尾时间，程序同样会进行检测，如果不满足要求，会进行插值处理。

针对本模型，选项（Setting）中除了需要指定基准时间（BaseTime）外，还有如下特有选项：

表 12.3-1 统计模型选项

参数名	含义	取值范围	默认值	备注
Freq_Day	重采样时间间隔（以天为单位）	'auto'或大于 0 实数	'auto'	用于 GM 模型
hidden_layer_size	隐藏层神经元数目	'auto'或大于 0 整数	'auto'	用于 BP 模型
activation	激活函数	'identity', 'logistic', 'tanh', 'relu'	'relu'	用于 BP 模型
kernel	核函数	'rbf'-径向基核函数/高斯核函数, 'poly'-多项式核函数, 'sigmod'-sigmod 核函数, 'linear'-线性核函数	'rbf'	用于 SVM 模型
C	惩罚系数	大于 0 实数	1.0	用于 SVM 模型
gamma	核系数	'scale', 'auto', 或大于 0 实数	'scale'	用于 SVM 模型，对多项式核函数无效
degree	多项式核函数阶次	大于 0 整数	3	用于 SVM 模型，对非多项式核函数无效
epsilon	SVR 参数	大于 0 实数	0.1	用于 SVM 模型

12.4 计算流程

12.4.1 模型训练

模型训练时，依次执行下列流程：

（1）调用通用预处理模块，按照因子设置进行处理，获得处理后原因量；

（2）检查数据是否时间间隔相等，如不相等，按照 Setting 中 Freq_Day 项设置，如果该选项为 auto，则计算平均时间间隔，对数据进行重新采样；如果该选项为大于 0 的实数，则按照该时间间隔（以天为单位）进行重新采样，确保数据时间间隔相等；

（3）调用 GM 模型进行训练；

（4）根据 Setting 设置构建 BP 模型，进行训练；

（5）根据 Setting 设置构建 SVM 模型，进行训练；

（6）根据三个模型的计算结果和实测结果，采用最优加权组合方法，计算三个模型的组合权重。

（7）根据组合权重将三个模型的结果进行组合，获得最终结果。

12.4.2 模型预测

模型预测时，依次执行下列流程：

（1）调用通用预处理模块，按照因子设置（已在训练时保存）进行处理，获得处理后原因量；

（2）检查数据是否时间间隔相等，如不相等，按照 Setting 中 Freq_Day 项设置，如果该选项为 auto，则计算平均时间间隔，对数据进行重新采样；如果该选项为大于 0 的实数，则按照该时间间隔（以天为单位）进行重新采样，确保数据时间间隔相等；另外检查预测数据的起始时间是否紧接训练数据的终止时间，如果不满足要求，则通过线性插值进行调整。

（3）分别调用 GM、BP、SVM 模型进行预测，获得计算结果；

（4）根据训练获得的最优组合权重将三个模型的结果进行组合，输出最终结果。

12.5 模型输出

返回结果为 JSON 格式，具体参数的含义见表 12.4-1。

表 12.5-1 GM-BP-SVM 组合模型输出信息

参数	类型	含义
yReal	数组	原始效应量数据，跟输入值相同
yCalc	数组（跟 yCalc 等长）	模型计算出的效应量数据
yComponent	对象	GM、BP、SVM 子模型计算结果
Summary	字符串	建模过程输出信息，包含护输出输入原因量和处理后原因量，总体精度，方差分析表，回归系数表等。
Evaluate	对象	模型总体精度数据，包含'R2'-决定系数，'RMSE'-标准误差，'R'-相关系数。
ModelFile	字符串	模型文件路径
StatusCode	整数	状态号，成功为 200，有错误为 500
StatusMessage	字符串	如果发生错误，存储错误提示信息，详细信息在 Summary 中

12.6 其他说明

（1）由于本组合模型中含有 GM 模型，而 GM 模型要求数据时间间隔相等，因此建议输入数据满足等时间间隔的条件。如不满足，程序会计算平均时间间隔，然后通过线性插值获得等时间间隔的数据，确保模型能够顺利运行，但结果的精度可能受到影响。另外在预测时，GM 模型要求预测的起始时间必须紧接训练的结尾时间，对此程序同样会进行检测，如果不满足要求，会进行插值处理。

（2）当原始时间时间间隔不相等时，本模型会通过插值算法进行处理，所以计算结果的时间会跟实测数据时间不对应，针对该问题，程序再最终输出时，会再次进行重采样处理，将计算结果的

插值到跟实测数据相同的时间点，方便用户对比。对用户而言，插值过程为程序内部步骤，不需要考虑，也不会感受到插值带来的时间不一致问题。

（3）本模型使用的 BP 子模型具有随机性，故相同的数据每次运行的结果都略有区别，这是神经网络的特点决定的；

（4）由于本模型不会输出类似统计模型的回归方程，因子处理后原因量的定义对大部分用户而言没有太大意义，上述信息仅会在返回结果的 Summary 中给出，供专业用户参考。

12.7 算例验证

12.7.1 算例说明

选用大坝 PL01XH01 测点 2013/10~2020/9 的监测数据，利用 2013~2019 年数据进行建模，然后预测 2020 年数据与实测数据进行比对，以验证模型的有效性。2013~2020 年环境量（原因量）和测点实测数据（效应量）的过程线如下图所示。

图 12.7-1 环境量和测点实测数据过程线

12.7.2 因子设置

选择因子设置如下：

Component（分量类型）	ItemType（测点类型）	Expression（因子计算表达式）	MaxOrder（最高多项式阶数）	备注
Head（水位分量）	Head_Up（上游水位）	x-75	2	
Head（水位分量）	Head_Down（下游水位）	x-75	2	
Temp（温度分量）	Temp_Air（气温）	x	1	
Temp（温度分量）	Temp_Air（气温）	Average(x,30)	1	30 天平均气温
Time（时效分量）	Time（时间）	ln(1+x/365)	1	x/365 转换时间单位为年

12.7.3 模型选项

模型选项均采用默认值，详见表 12.3-1。

12.7.4 总体结果

利用 2013~2019 年数据进行建模，然后预测 2020 年数据与实测数据进行比对，总体结果如下图所示。图中上半部分给出了模型拟合/预测值与实测值过程线，下半部分给出了各子模型（GM、BP、SVM）过程线。

下面给出训练和预测结果的具体数据。

12.7.5 训练结果

使用 2013~2019 年数据进行建模，结果如下：

（1）总体建模结果

模型	相关系数	决定系数	标准误差	组合权重
组合模型	0.969469	0.939780	0.138129	
GM 子模型	0.764295	-0.049683	0.576690	0.031629
BP 子模型	0.865474	0.747917	0.282609	-0.124339
SVM 子模型	0.968476	0.936851	0.141447	1.092711

从上表可以看出，GM 模型的拟合精度相对较低，BP 模型居中，SVM 模型最好，组合模型采用最优加权组合法，合理确定三个模型的权重，最终 SVM 模型权重最高，BP 次之，GM 最低，相对关系合理。组合模型的拟合效果优于三个子模型，表明最优加权组合方法是有效地。

（2）详细结果

效应量实测值、各模型拟合值和误差结果见下表。

12.7.6 预测结果

（1）总体建模结果

模型	相关系数	决定系数	标准误差
组合模型	0.920269	0.840580	0.172997
GM 子模型	0.606281	-0.835546	0.587017
BP 子模型	0.933073	0.784223	0.201266
SVM 子模型	0.922358	0.844110	0.171071

从上表可以看出，GM 模型的预测精度相对较低，BP 模型居中，SVM 模型最好，最优加权组合模型的预测效果与 SVM 模型一致。

（2）详细结果

效应量实测值、各模型拟合值和误差结果见下表。

12.8 小结

（1）由于实际大坝工作条件复杂，影响因素众多，单一的分析模型在稳定性上有所不足，本模型采用最优加权组合方法，将灰色系统模型（GM）、神经网络模型（BP）和支持向量机模型（SVM）有机结合，取长补短，有助于获得更优和更稳定的建模和预测效果。

（2）本模型算例建模和预测结果与实测结果规律高度一致，模型训练结果的总体精度指标较好，表明模型是有效的。从结果来看，GM 子模型的精度相对较低，SVM 模型精度最高，组合模型的效果不低于 SVM 模型。但上述结果仅是一个算例的结果，对于不同的数据，不同的选项设置，GM、BP、SVM 三个模型的效果可能有所变化，最优加权组合方法取长补短，采用最优的权重组合，确保最终计算结果的可靠性和稳定性。当用户建模经验不足，模型因子和选项设置不尽合理时，最优加权组合模型较单一模型具有更强的鲁棒性和智能性，有助于获得更优的计算结果。

（3）本模型针对各子模型提供了较为丰富的设置参数，专业用户可以根据具体情况对模型进行深入的调试。

（4）由于本组合模型中含有 GM 模型，而 GM 模型要求数据时间间隔相等，同时预测数据必须紧接训练数据，因此程序内部会进行检测和插值重采样处理，不需要用户额外考虑，进一步提高了模型的智能性。

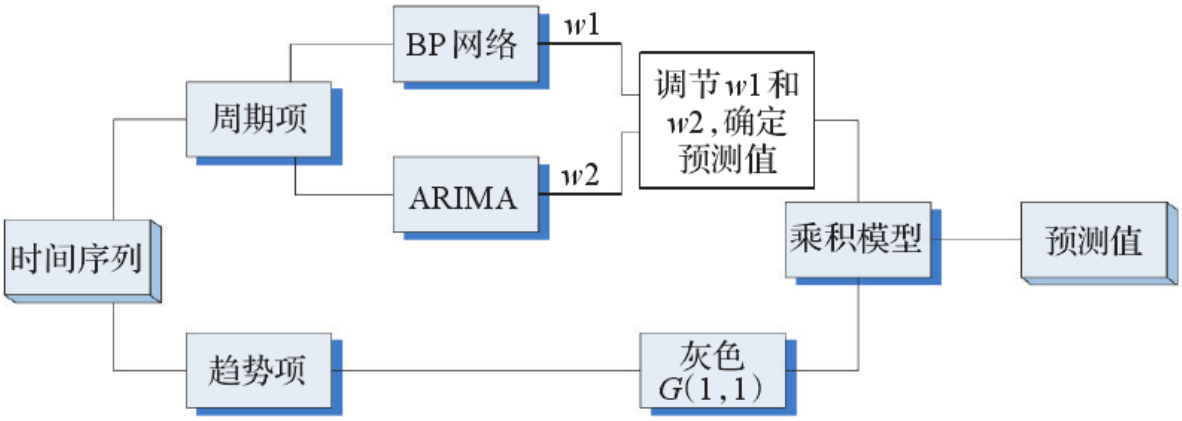
12.9 参考文献

13. 统计-BP-SVM 智能组合模型

14. 趋势变化和周期波动组合模型

由于大坝监测时间序列呈现趋势变动性和周期波动性特征，对这种二重时间序列预测提出了很多方法，其中最常见的是自回归滑动平均（ARIMA）模型，而该模型要求时间序列数据经过差分后具有平稳性；BP 神经网络也广泛应用于该时间序列的预测，但它常常会忽略某些巨大噪音或非平稳数据；而灰色 G（1，1）模型，但它仅能较好地拟合时间序列的趋势性部分，而对于周期波动性，其预测精度则明显降低。显然，若用这些单一的模型对复杂的二重时间序列进行预测，难以取得理想效果。将灰色 G（1，1）模型、BP 模型、ARIMA 模型进行智能组合，使之符合二重趋势时间序列的特征。具体算法如下：

首先对二重趋势时间序列进行分解，得到趋势变动项和周期波动项，用灰色 GM（1，1）对趋势项预测，再用 BP 神经网络和 ARIMA 的组合模型对周期项预测，最后用乘积模型合成这两部分结果，得到二重趋势时间序列的最终预测值。其智能算法如下图所示。



几个关键步骤如下：

（1）二重趋势分解

对二重趋势特征的时间序列数据进行分析应用时，通常可分解成趋势变动项和周期波动项，典型的分解方法是乘积模型：

$$X(t)=P(t)\cdot T(t)$$

其中： $X(t)$为观测值，$P(t)$为观测值的趋势变动项，$T(t)$为观测值的周期波动项。

设序列$X(1)$，$X(2)$，...，$X(h)$，其中 h 为序列的长度。用中位移动平均法，可以提取不含周期波动的趋势项：

$$P(t)=\frac{1}{12}[\sum_{i=-5}^{5}X(t+i)+\frac{1}{2}(X(t+6)+X(t-6)],t=7,8,...h-6$$

式中是以 12 为周期，以t为中心，2 阶对称滑动平均数字滤波，P（t）经过滤波后不再含有周期波动项。实际使用时，可以按照数据的采用间隔等信息，灵活指定周期数。

对应的周期波动项为

$$T(t)=\frac{X(t)}{P(t)}$$

这样就把观测数据的时间序列分解为趋势项和周期项，可分别对它们分析、建模和预测。

（2）趋势变动项预测

对于分离后的趋势变动项，使用 GM（1,1）模型求得拟合曲线并预测。

（3）周期波动项预测

周期波动项具有非常复杂的非线性结构，对其准确预测较难。BP 神经网络模型是目前比较成熟的算法，在函数逼近和数据拟合方面具有很大优越性，但它常会忽略一些大的噪音数据。而自回归滑动平均（ARIMA）模型对噪音数据具有很强的预测能力。由此，结合这两个模型的优势，分别用 BP 神经网络模型和 ARIMA 模型对周期项波动预测，根据权重优化模型，建立周期波动项的组合预测模型。

ARIMA(p,d,q)模型，即对输入数据进行 d 阶差分，使之平稳，p 为自回归项数，q 为滑动平均项数。ARIMA(p,d,q)也就是对数据进行 d 阶差分后，输入到 ARMA(p,q)模型中。其中 ARMA 模型建立过程可见其他章节。

参照前文方法分别建立BP神经网络模型和ARIMA模型。周期的组合模型建立方法如下：

设$\hat{y}_{BP}$（k）是BP神经网络对周期波动项的预测值，$\hat{y}_{ARIMA}$（k）是 ARIMA 模型对周期波动项的预测值，则周期项的组合预测值$\hat{y}$（k）为：

$$\hat{y}(k)=w_{bp}\hat{y}_{bp}(k)+w_{ARIMA}\hat{y}_{ARIMA}(k)$$

式中w_{BP}、w_{ARIMA}为每种预测值的权重，目标函数取误差平方和最小，若有l个实际观测值，则周期波动项的组合模型权重的优化模型如下：

$$\min\sum_{k=1}^{l}\left|y(k)-\hat{y}(k)\right|^2$$

$$s.t.\ w_{bp}+w_{ARIMA}=1$$

其中，权重w_{BP}、w_{ARIMA}的确定可以通过最优加权组合法得到。

15. EEMD-SVM-ARMA 组合模型