基于 SARIMA 预测的黄河水沙监测分析模型

摘要

摘要黄河, 作为中华民族的母亲河, 是孕育中华民族千百年历史文化源泉。然而近几年经济快速发展带来的环境破坏导致黄河水沙问题日益严峻。因此研究黄河水沙通量的变化规律对于环境治理、协调人地关系等具有重要意义。

对于问题一, 分析可得本问共由两个部分组成。一是研究含沙量与时间、水位和水流量的关系，考虑到附件一提供数据存在缺失值，所以首先对附件 1 数据进行数据预处理，通过线性插值处理缺失值，后计算含沙量分别与水位、流量两者的皮尔逊相关系数确定其是否具有线性关系，计算结果分别为 0.6456 与 0.6469 存在线性关系。随后对其进行线性拟合与对数拟合求解拟合函数，最后对含沙量与时间关系进行三角函数拟合，求解拟合结果，获得含沙量与三者关系。二是估算每年的年总水流量和年总排沙量。所以我们构建年总水流量和年总排沙量的计算公式，并直接求得每年的结果。计算完成后我们进行拟合检验分别对含沙量与时间、水位、流量之间关系所求的拟合函数进行拟合优度与 T 的相伴概率的计算，根据拟和检验结果进行模型准确性与模型性能评价，保证模型具有较高的精度与性能。

对于问题二, 本问中要求分析进六年水文站水沙通量的突变性、季节性和周期性等变化规律。鉴于附件 1 提供数据存在缺失值且数据量较大，不利于对其规律进行分析，因此首先对其缺失值进行线性填充，然后进行特征提取，对原始数据进行基本的描述性分析。对于周期性，决定利用傅里叶变化展开时间序列，分析频谱图中的峰值得到周期，确定其周期均为一年。对于季节性，考虑利用了各个月的平均水沙通量以及季节因子来表明各个月之间存在的差别，。对于突变性，考虑利用 M-K 检验方法对其进行处理，输出 $u_f(k)$ 和 $u_b(k)$ 图形以得到较明显的突变点，确定水通量与沙通量突变所在时间分别为 18 年 4 月与 18 年 5 月，完成对水沙通量的变化规律分析。

对于问题三, 分析可得本问共有两部分组成。一是根据水文站水沙通量的变化规律，预测水文站未来两年水沙通量的变化趋势，所以为确定本组时间序列模型是否未来存在相同趋势、反转趋势或随机趋势，模型决定应用 R/S 分析方法进行检验，计算得水月通量与沙月通量的 Huist 指数分别为 0.83857 与 0.76523 其 R/S 值分别为 19.34843 与 15.87536。针对水沙通量的预测，考虑到时间序列具有季节性等特性，因此应采用 SARIMA 预测模型，以前五年数据为训练集，2021 年数据为验证集，并对 2022-2023 年数据进行预测分析最后将结果输出于附件 2。二是制定未来两年最优的采样检测方案，通过利用小波分析法对得到的数据分析，确定旱涝灾害时期，合理分配监测次数与具体时间。

对于问题四, 本问要求根据水文站水沙通量与河底高程变化情况，分析调水调沙的实际效果，并对不进行调水调沙十年后水文站河底高程进行预测。通过对提供的附件进行研究，发现附件 2 勘测的距离并非统一长度，所以模型选取了共有的一段进行高程分

析，求得了平均高程分析调水调沙作用。对于水通量和沙通量，模型通过同时分析六年数据，来分析调水调沙的效果。最终研究未调水调沙时段的河底高程变化得到 10 年之后的高程变化。

最后，本模型从函数推导与理论分析两个角度求解了水文站黄河水的含沙量与时间、水位、水流量的关系，估计了近六年水文站年总水量和排沙量；分析了近六年水文站突变型、周期性和季节性等变化规律；实现对水沙通量的预测，制定了采样检测方案，最后分析了调水调沙实际效果，并对未调水调沙的水文站河底高程进行预测。建模较为成功。

关键字：拟合　特性分析　SARIMA 预测

一、 问题重述

1.1 问题背景

黄河，作为中华民族的母亲河，千百年来承载着中华深厚的历史文化，是中国北方地区生生不息的力量源泉。近年来随着人类生活对环境压力的不断增加，黄河水沙问题日益加剧，不断困扰着流域内的生态环境、经济发展以及人民生活的安全。因此本小组依托小浪底水库下游某重要水文站近六年的详细监测数据，包括水位、水流量、含沙量等关键指标，以及该水文站黄河断面的测量数据，揭示黄河水沙通量的变化特征。希望通过这一研究能够为黄河流域的综合治理和可持续发展提供有力的数据支撑和科学依据，助力实现“绿水青山就是金山银山”的发展理念，推动黄河流域生态保护和高质量发展。

1.2 问题要求

问题 1 首先根据附件 1 提供的监测数据，分析黄河水的含沙量与时间、水位、水流量之间的关系, 然后基于提供的数据，计算每年黄河通过该水文站的总水流量和总排沙量。

问题 2 首先识别水沙通量在时间序列上是否存在显著的突变点，即突然增加或减少的情况。然后分析水沙通量是否具有季节性变化模式和长期周期性。进而总结水沙通量的主要变化规律。

问题 3 首先根据问题二中得出的水沙通量变化规律，利用时间序列预测模型预测未来两年的水沙通量。后基于预测结果，设计最优的采样监测方案，包括确定采样监测的次数、具体时间点等，使其既能及时掌握水沙通量的动态变化情况，又能最大程度地减少监测成本资源。

问题 4 首先根据附件 2 中的黄河断面测量数据，结合每年 6-7 月小浪底水库进行“调水调沙”期间的水沙通量数据，分析这一操作对下游水文站河段的实际效果。之后基于当前的水沙通量变化规律和河底高程数据，预测如果不进行“调水调沙”，10 年以后该水文站的河底高程将如何变化。

二、 问题分析

2.1 问题一分析

对于问题一，分析可得共由两个部分组成。一是研究含沙量与时间、水位和水流量的关系，二是估算每年的年总水流量和年总排沙量。所以应首先观察了其各个变量之间

的散点关系图，确定其关系并进行求解，并进行参数有效性分析，确保其合理性，提高模型准确性。之后构建年总水流量和年总排沙量的计算公式，并求得每年的结果。v

2.2 问题二分析

对于问题二，本问中要求分析进六年水文站水沙通量的突变性、季节性和周期性等变化规律。对于季节性，考虑利用了各个月的水沙通量以及季节因子来表明各个月之间存在的差别，对于周期性，决定利用傅里叶变化展开时间序列，分析得到周期。对于突变性，考虑利用 M-K 检验方法对其进行处理，以得到较明显的突变点，完成对水沙通量的变化规律分析。

2.3 问题三分析

对于问题三，首先为确定本组时间序列模型是否未来存在相同趋势、反转趋势或随机趋势，考虑应用 R/S 分析方法进行检验。针对水沙通量的预测，考虑到时间序列具有季节性等特性，因此应采用 SARIMA 预测模型，以前五年数据为训练集，2021 年数据为验证集，并对 2022-2023 年数据进行预测分析。最终对得到的数据利用小波分析法分析旱涝灾害时期，合理分配监测次数。

2.4 问题四分析

对于问题四，通过观察提供的附件，发现附件 2 勘测的距离并非统一长度，所以模型选取了共有的一段进行高程分析，求得了平均高程分析调水调沙作用。对于水通量和沙通量，模型通过同时分析六年数据，来分析调水调沙的效果。最终研究未调水调沙时段的河底高程变化得到 10 年之后的高程变化。

三、 模型假设

为简化问题，本文做出以下假设：

- 假设 1 附件提供的数据是准确无误且完整的
- 假设 2 输出分析中，水位、水流量与含沙量之间存在线性关系。
- 假设 3 忽略除已提供数据外的其他外部因素，如降雨量等的影响。
- 假设 4 “水位”和“河底高程”均以“1985 国家高程基准”（海拔 72.26m）为基准面
- 假设 5 季节性影响可以通过历史数据中的周期性模式来捕捉。

四、 符号说明

符号	说明	单位
m	质量	kg
V	体积	m^3

五、 问题一的模型的建立和求解

5.1 模型建立

5.1.1 数据预处理

鉴于附件 1 所提供数据为小浪底水库下游黄河某水文站实测数据，数据量较大且数据存在偶然性与突发性，因此需要对数据进行预处理，以过滤偶然信息，提取矩阵信息，提高模型的准确性与精确度同时便于模型的建立与求解。

1. **缺失值数据：**
   分析可知在附件 1 中所提供的黄河含沙量数据为测量当天多个时间段的随机数据，不连续且时间段偶然性较强，为提高模型准确性避免偶然性，本模型提取每天 8 点的含沙量数据进行处理，该时间段不存在数据则当天为缺失值。
2. **缺失值处理:**
   含沙量数据处理后，发现存在明显的数据缺失，考虑含沙量数据为一系列离散观测值，所以本模型考虑通过线性插值估算缺失数据。给定一系列含沙量数据 C_{t_i} 和对应时间点 t，按照时间序列新一个缺失值前一个数据必为非缺失数据，因此通过缺失值前一个序列数据与后第一个非缺失值数据进行线性拟合，以确定缺失值数据，数学公式如下:

$$C_t = \frac{d}{d+1} C_{t-1} + \frac{1}{d+1} C_{t+1} \tag{1}$$

   其中: C_{t-1} 是时间点 t_{t-1} 上的含沙量观测值，C_{t+d} 是时间点 t_{t+d} 上的含沙量观测值，t_{i-1} 和 t_i 是已知含沙量观测值的时间点，t 是要估算含沙量的时间点。同样若含沙量时间序列中存在多个缺失值，依然可以通过迭代应用上述公式对缺失值进行估算，以确保散点图数据完整提高模型的准确度。

5.1.2 皮尔逊相关性的关联分析

皮尔逊相关系数的相关性分析通常用于探索两个变量之间是否存在线性相关性，其绝对值大小可以反映反映了变量间相关性的强度（越接近 1 其相关性越强），其具体实

现如下：给定两个变量 X 和 Y 的样本 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$，皮尔逊相关系数 r 可以通过以下公式计算：

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

其中: $\bar{x}$ 是变量 X 的平均值，$\bar{y}$ 是变量 Y 的平均值，n 是样本数量。

在本问中模型建立的主要目的是分析含沙量与流量、水位之间的关系，经散点图观察可初步认为其存在线性关系，因此应用皮尔逊相关系数进行验证与求解，其中分别将流量和含沙量、水位和含沙量作为 X、Y 进行计算得出其皮尔逊相关系数 r_1、r_2 分别为 r_1=0.6469、r_2=0.6356 线性相关性较强，而在流量小于 2000 立方米时 r_1=0.6729，在含沙量小于 20 千克每立方米时 r_2=0.7178 线性相关性更强，所以可以建立线性拟合确定两者关系，关联性分析较为成功。

5.1.3 构建回归模型

在本模型中通过对数据分析，我们已知数据具有较强的线性相关性，所以首先希望获得对数据关系拟合的最直观模型，方便对数据进一步处理与关系的进一步确定，又考虑到数据跨度较大，所以添加对数拟合实现更大范围的捕捉数据的变化规律，在含沙量与时间关系拟合中，考虑到时间序列的周期性，本模型选择应用三角函数进行拟合，最后通过对拟合优度与 T 检验的相伴概率进行分析，确定最优拟合模型，并检验其精确度与准确性，提高模型的鲁棒性。

1. 线性拟合

   线性拟合是最常见的数据拟合方法之一，适用于当变量间的关系大致呈现线性时的情况。其通过最小二乘法找到一条最佳拟合直线，使得所有数据点到直线的距离平方和最小，具体实现如下：对于一组数据点 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$，最小二乘法的线性拟合可以表示为：

   $$y = ax + b \tag{3}$$

   其中，斜率 a 和截距 b 可以通过以下公式计算得出：

   $$a = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$
   $$b = \frac{\sum y - a\sum x}{n}$$

   其中：n 表示数据点的数量，$\sum xy$ 表示所有 $x_i y_i$ 的总和，$\sum x$ 和 $\sum y$ 分别表示所有 x_i 和 y_i 的总和 $\sum x^2$ 表示所有 x_i^2 的总和。

2. 对数拟合

对数拟合通常用于数据在对数尺度上呈现出线性关系的情况。其主要通过对数据取对数，然后使用最小二乘法拟合一条直线来实现。这种拟合方法适用于数据在较大范围内变化且呈幂律分布的情况，具体实现如下：假设原始数据为 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$，首先对数据取对数得到 $(\log(x_1), \log(y_1)), (\log(x_2), \log(y_2)), \ldots, (\log(x_n), \log(y_n))$。后使用最小二乘法来拟合对数坐标系下的数据点 $(\log(x_i), \log(y_i))$。假设对数坐标系下的直线方程为：

$$\log(y) = a \log(x) + b \tag{4}$$

转换回原始坐标系，我们得到幂律关系：

$$y = x^a \cdot e^b$$

其中，a 和 b 可以通过对数坐标系下的最小二乘法求得。

3. 三角函数拟合

三角函数拟合是指使用三角函数（如正弦和余弦函数）来近似一组数据点的过程。其主要通过确定周期后，以最小二乘法实现，这种拟合通常用于周期性数据的分析，对于时间序列的处理有较为优异的性能，具体实现如下：假设目标函数 $f(x)$ 是一个正弦函数形式：

$$f(x) = A \sin(\omega x + \phi) + B,$$

其中：A 是振幅，ω 是角频率，ϕ 是相位偏移，B 是基线偏移。假设存在一组数据点 (x_i, y_i)，$i = 1, 2, \ldots, n$。下面需要最小化误差平方和 E：

$$E(A, \phi, B) = \sum_{i=1}^{n} \left(y_i - (A \sin(\omega x_i + \phi) + B)\right)^2.$$

为了估计参数 $A, \phi,$ 和 B，需要求解上述方程中的偏导数并令其等于零。即求解下列方程组：

$$\begin{cases} \dfrac{\partial E}{\partial A} = -2 \displaystyle\sum_{i=1}^{n} \left(y_i - (A \sin(\omega x_i + \phi) + B)\right) \sin(\omega x_i + \phi) = 0, \\ \dfrac{\partial E}{\partial \phi} = -2 \displaystyle\sum_{i=1}^{n} \left(y_i - (A \sin(\omega x_i + \phi) + B)\right) \cos(\omega x_i + \phi) \cdot A\omega = 0, \\ \dfrac{\partial E}{\partial B} = -2 \displaystyle\sum_{i=1}^{n} \left(y_i - (A \sin(\omega x_i + \phi) + B)\right) = 0. \end{cases}$$

又由于这些方程是非线性的，通常不能直接解析求解，而是采用数值方法来求解。因此选的梯度下降法或牛顿迭代法等求解，这里不再详细赘述。通过应用三种拟合方式对数据进行处理，实现了对数据间关系求解的最优化处理，同时从不同角度出发对数据进行对比，有利于提高模型鲁棒性，优化模型性能。

5.1.4 拟合检验

1. 拟合优度

   拟合优度是评估模型拟合数据好坏程度的一个重要指标, 本模型中考虑到拟合依靠多种模型建立，因此选择普遍性最强的决定系数 R^2 作为度量标准。决定系数 R^2 主要用于衡量模型对数据的解释能力，在回归分析中，用于表示模型解释的变异量占总变异量的比例。决定系数 R^2 值介于 0 到 1 之间，值越接近 1 表示模型拟合数据的程度越高。其具体实现如下：对于一组数据点 (x_i, y_i)，其中 $i = 1, 2, \ldots, n$，以及对应的模型预测值 $\hat{y}_i$，决定系数 R^2 可由以下公式求得：

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

   其中: - y_i 是第 i 个观测值，- $\hat{y}_i$ 是第 i 个观测值的预测值，- $\bar{y}$ 是所有观测值的平均值，- n 是数据点的数量。通过对决定系数 R^2 的分析可以更为直观的确定模型的数据的拟合情况，方便对模型的优化与选择，大大提高了模型性能。

2. T 的相伴概率

   T 的相伴概率即 t-检验的 p 值（p-value），是评估统计显著性的关键指标之一。主要用于用于比较两个样本均值之间的差异是否显著，在本模型中用于评估线性回归模型斜率系数的显著性。指在假设斜率为 0（即变量之间不存在线性关系）的情况下，观测到当前斜率或更极端斜率的概率。其具体实现如下：线性回归模型的一般形式为：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

   其中: y_i 是第 i 个样本的响应变量；x_i 是第 i 个样本的解释变量；β_0 是模型的截距；β_1 是模型的斜率；ε_i 是随机误差项，通常假设其服从正态分布。最小二乘估计的目标是最小化残差平方和（RSS），即：

$$RSS = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

   最小化这个表达式会给出斜率 β_1 和截距 β_0 的估计值。斜率 β_1 的估计值 $\hat{\beta}_1$ 的 t-统计量定义为：

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)}$$

   其中: - $\hat{\beta}_1$ 是斜率的估计值；- β_1 是假设的斜率值（在零假设下，$\beta_1 = 0$）；- $\text{SE}(\hat{\beta}_1)$ 是 $\hat{\beta}_1$ 的标准误差。t-统计量 t 的 p 值可以通过 t-分布来计算。在零假设 $H_0 : \beta_1 = 0$ 下如果 p 值小于预先设定的阈值 0.05，则认为变量之间的线性关系是统计上显著的，可以拒绝斜率为 0 的原假设，认为斜率 β_1 显著不为 0 否则则相反。

5.1.5 求解年总水流量与年总排砂量

根据附件 1 所提供的六年内每天的河流含沙量与流量数据，可直接对年总水流量与年总排砂量进行求解，具体实现如下：

1. 年总水流量
   设求第 i 年的年总水流量。

$$Q_i = 24 * 3600 * (\text{这一年每天的流量均值求和})$$

2. 年总含沙量 设求第 i 年的年总含沙量。

$$Q_2 = 24 * 3600 * (\text{这一年每天的流量均值再乘这一天 8 小时含沙量的和})$$

5.2 模型求解

通过对问题一的分析确定最后模型求解主要为数据预处理、数据拟合、拟合检验三步具体实现如下：

Step1：处理附件 1 数据，提取未缺失数据，处理缺失值，最后汇总六年数据于一个矩阵数据，方便计算机进行处理。

Step2：首先计算含沙量分别与水位、流量两者的皮尔逊相关系数，确定其是否具有线性关系，若具有线性关系对其进行线性拟合与对数拟合求解拟合函数，最后对含沙量与时间关系进行三角函数拟合，求解拟合结果。

Step3：进行拟合检验，分别对含沙量与时间、水位、流量之间关系所求的拟合函数进行拟合优度与 T 的相伴概率的计算，根据拟和检验结果进行模型准确性与模型性能评价。

Step4：最后根据附件 1 直接求解近六年该水文站的年总水流量和年总排沙量。

5.3 求解结果

1. 皮尔逊相关系数
   根据散点图初步分析可得含沙量与流量与水位的关系具有较好的线性关系，而含沙量与时间的关系更多的是周期性关系，因此只对前两部分进行皮尔逊相关系数计算，为突出其相关性关系计算过程中分别对两部分进行了一次范围设定获得更高的线性关系，过滤一部分突出数据，使最终结果更具有普遍性与鲁棒性，其具体结果如下表：经分析含沙量与流量和水位之间都具有较好的正相关的线性关系，与观察结果一致。
2. 拟合函数

表 1　皮尔逊相关系数

含沙量与流量	0.64687
含沙量与流量（流量小于 2000m³）	0.67293
含沙量与水位	0.63563
含沙量与水位（含沙量少于 20 kg/m^3 ）	0.71776

通过对散点图观察与 step1 的分析，决定对含沙量与流量和含沙量与水位的关系进行线性拟合与对数拟合，对含沙量与时间的关系进行三角函数的拟合，以保证求得数据关系的最优拟合函数，其具体结果如下表：

表 2　含沙量与流量的关系拟合方程

拟合直线方程	y=0.00385x-0.58768
拟合对数方程	ln（含沙量）=1.15063*ln（流量）-6.94861

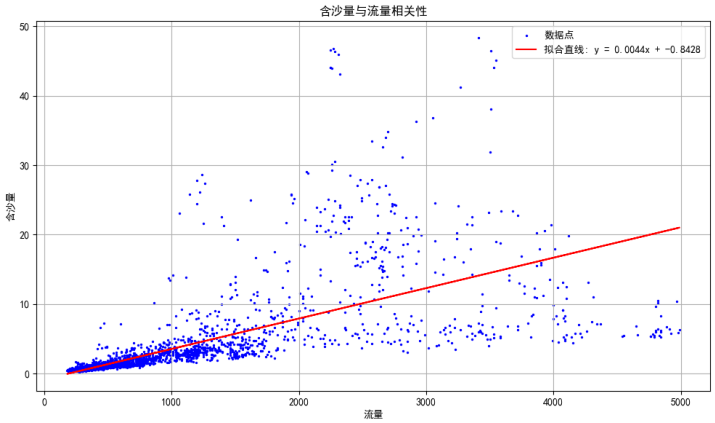

图 1　含沙量与流量关系图

表 3　含沙量与水位的关系拟合方程

拟合直线方程	Y=2.82290x-118.58640
拟合对数方程	ln（含沙量）=40.11907*ln（水位）-150.27277

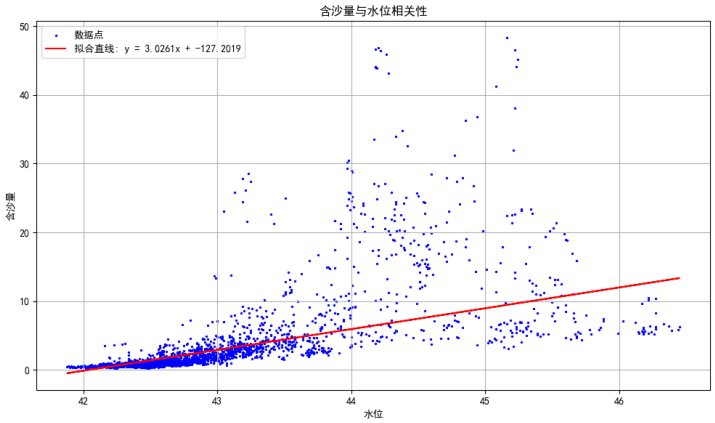

图 2　含沙量与水位关系图

表 4　含沙量与时间的关系拟合方程

前两年	y=-0.3294*cos(2*pi/365*day+1.251)+0.934
后四年	y=-4.049*cos(2*pi/365*day+12.09)+4.674

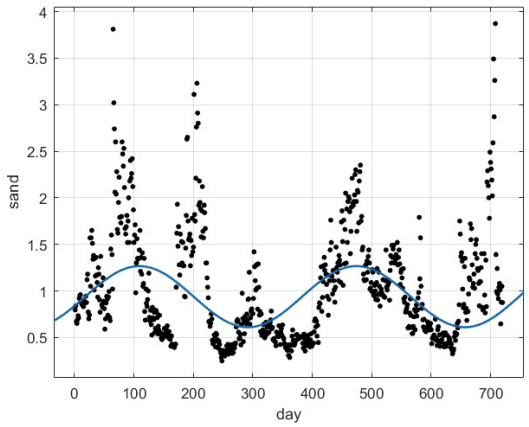

图 3　含沙量与前两年时间关系图

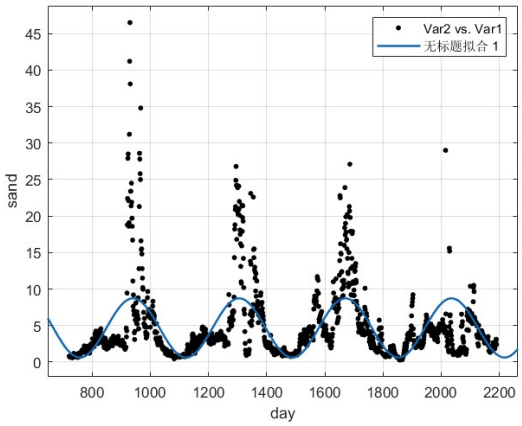

图 4　含沙量与后四年时间关系图

观察散点图与拟合曲线不难发现其重叠程度较高，因此可以初步判断建模较为成功，下面进行拟合检验以进一步确定模型精度与性能。

3. 拟合检验

通过计算得出含沙量与水位、流量和时间的关系方程，为判断其准确性与鲁棒性，对其进行拟合检验，求解其拟合优度 R^2 以确保所求函数能正确反应两者之间的关系，提高模型的准确性与适用性。

表 5　拟合检验

	拟合优度（R^2）	T 检验的相伴概率
含沙量与流量	0.802	0
含沙量与水位	0.627	0
含沙量与时间（前两年）	0.396	
含沙量与时间（后四年）	0.298	

以可以判断可以本模型可以较为准确的表达出含沙量与流量、水位、时间之间的关系，模型建立较为成功，具有较强的准确性与鲁棒性。

4. 年总水流量和年总排沙量

根据附件 1 提供数据直接计算近六年年总水流量和年总排沙量，结果如下表所示：

表 6　年总水流量与年总排沙量

	年总水流量	年总排沙量
第一年	14345164800	18238245508
第二年	15365721600	19068119604
第三年	38998828800	294642170776.8
第四年	38789683200	305780228611.20
第五年	43770369600	3.53254E+11
第六年	47425471200	228220490070

六、 问题二的模型的建立和求解

6.1 模型建立

6.1.1 数据处理

考虑到附件 1 所提供数据存在缺失值，因此对其进行预处理对缺失值进行线性填充，以保证数据的连续性完整性，相关处理见问题一模型建立。随后通过对附件 1 的分析，不难发现对其中数以千计的数据直接进行处理是不现实的，因此为更好的理解和分析数据，应对数据进行特征提取从原始数据中提取有用的统计信息和其他有意义的特性进行分析，所以对数据以月为单位进行了如下的特征提取：

1. 均值: 表示数据集的中心趋势。公式表达如下：

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

2. 标准差: 表示数据的离散程度。公式表达如下：

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

3. 最小值：数据集中最小的观测值。公式表达如下：

$$x_{\min} = \min(x_1, x_2, \ldots, x_N)$$

4. 最大值：数据集中最大的观测值。公式表达如下：

$$x_{\max} = \max(x_1, x_2, \ldots, x_N)$$

5. 偏度：表示数据分布的不对称性。正偏度表示长尾在右侧，负偏度表示长尾在左侧，公式表达如下：

$$g_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right)^{3/2}}$$

6. 峰度：描述分布的峰值形状。高高正峰度表示尖峰分布，低负峰度表示扁平分布。公式表达如下：

$$g_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right)^2} - 3$$

6.1.2 季节性分析

为更为具体的开展水文站水沙通量的季节性分析，引入季节因子进行表示，其主要用来衡量特定时间段内数据相对于整个时间段平均值的偏差，可以帮助直观理解数据随时间变化的季节性模式。季节因子通常通过比较特定时间段（如一个月）的数据均值与整个时间段的数据均值来计算，其具体实现如下：假设我们有一组时间序列数据 x_t，其中 t 表示时间点（例如，月份）。我们可以按月计算数据的平均值，然后计算季节因子 S_m，具体实现如下：

1. 首先计算每个月的数据均值 $\bar{x}_m$，其中 m 表示月份。在数据处理中已进行计算。
2. 计算整个时间段的数据均值 $\bar{x}$。数学公式表示为：

$$\bar{x} = \frac{1}{M \cdot N}\sum_{m=1}^{M}\sum_{t \in \text{Month } m} x_t$$

3. 季节因子 S_m 为每个月的数据均值除以整个时间段的数据均值。用数学公式表示为：

$$S_m = \frac{\bar{x}_m}{\bar{x}}$$

6.1.3 周期性分析

鉴于附件 1 所提供时间序列数据在周期性模式上并非十分明显，所以考虑将数据转换到频率域，这时周期性模式将会表现为频谱中的峰值，这使得识别周期变得更加直观和简单，同时频率域提供了另一种视角来观察数据，提供了另一种特征提取的思路，利于更具体的分析数据变化规律。

本模型应用傅里叶变换将时间序列数据从时间域转换到频率域，进而识别出数据中存在的周期成分。这里对于提供的离散数据，对其进行离散傅里叶变换 (DFT) 处理，具体实现如下：对于长度为 N 的离散时间序列 x_n，其离散傅里叶变换 X_k 可以表示为：

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i k n}{N}}, \quad k = 0, 1, \ldots, N-1$$

数据转换之后应对傅里叶变换结果进行分析，以确定数据中的主要周期，主要通过对频率分量的能量进行分析得到，具体实现如下：傅里叶变换结果中每个频率分量的能量 E_k 可以用该分量的模的平方来表示：

$$E_k = |X_k|^2$$

其主要周期通过寻找能量 E_k 的峰值来确定。这些峰值对应于时间序列中的主要周期，通过对频谱图可视化，并依据时间序列进行划分可确定数据周期性。同时考虑到最终输出的频谱图可能会受到直流分量的影响，决定将影响明显的频谱图做去均值处理，具体实现如下：

6.1.4 突变性分析

突变性分析的目的是检测时间序列中是否存在显著的变化点或趋势并确定其突变型规律。通过分析确定数据不符合特定的概率分布，且在时间序列中存在长期趋势，并且原始数据存在缺失值对其进行了线性填充，所以决定应用 MK（Mann-Kendall）检验算法进行检验，Mann-Kendall 检验是一种常用的非参数统计测试，用于检测时间序列中是否存在趋势，具体实现如下：

1. 计算秩次：
   将时间序列数据 $x_1, x_2, \ldots, x_n$ 排序，计算每个数据点的秩次 r_i。
2. 计算秩次差：
   对于每一对时间点 (x_i, x_j)（其中 $i < j$），定义 s_{ij}：

$$s_{ij} = \begin{cases} 1 & \text{if } x_j > x_i \\ 0 & \text{if } x_j = x_i \\ -1 & \text{if } x_j < x_i \end{cases}$$

3. 计算统计量 S：

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s_{ij}$$

4. 计算方差 $Var(S)$：

$$Var(S) = \frac{1}{18}[n(n-1)(2n+5) - \sum_{p=1}^{g} t_p(t_p-1)(2t_p+5)]$$

   其中 g 是不同秩次的个数，t_p 是具有相同秩次的数据点的数目。
5. 计算标准化统计量 Z：
   如果 $S > 0$：

$$Z = \frac{S-1}{\sqrt{Var(S)}}$$

   如果 $S = 0$：

$$Z = 0$$

   如果 $S < 0$：

$$Z = \frac{S+1}{\sqrt{Var(S)}}$$

6. 确定显著性水平：如果 $|Z| > Z_{\alpha/2}$，则拒绝零假设（即没有趋势），其中 $Z_{\alpha/2}$ 是标准正态分布的 $\alpha/2$ 分位数。

M-K 检验的最终结果会给出一个 p 值，如果 p 值小于某个显著性水平（在本模型中预先设定其为 0.05），则可以认为存在显著的趋势变化，否则则无。而为更清晰的测定时间序列数据中的趋势变化点。引入 $u_f(k)$ 和 $u_b(k)$ 图形帮助识别数据中是否存在突变点，并估计突变发生的时间位置。其中 $u_f(k)$ 表示从时间序列的开始到第 k 个观测值的累积差异，$u_b(k)$ 表示从时间序列的结束到第 k 个观测值的累积差异。其计算方式如下：

对于时间序列数据 $x_1, x_2, \ldots, x_n$：

1. 计算 $u_f(k)$：

$$u_f(k) = \frac{S(k)}{\sqrt{Var[S(k)]}} - \frac{1}{2}$$

其中 $S(k)$ 是前 k 个数据点的 Mann-Kendall 统计量，$Var[S(k)]$ 是 $S(k)$ 的方差。

2. 计算 $u_b(k)$：

$$u_b(k) = \frac{S(n-k+1)}{\sqrt{Var[S(n-k+1)]}} - \frac{1}{2}$$

其中 $S(n-k+1)$ 是从第 k 个数据点到最后一个数据点的 Mann-Kendall 统计量。

6.2 模型求解

通过对问题二进行分析确定模型通过数据处理、周期性分析、季节性分析、突变性分析四步完成，具体实现如下：**Step1：**数据处理：首先检查附件 1 提供数据，对缺失值进行线性填充使数据保存完整连续，后提取数据特征 (平均值、标准差、最大值、最小值、峰值和偏度）对原始数据进行基本的描述性统计分析。

Step2：周期性分析：首先使用傅里叶变换（FFT）分析时间序列数据，以识别潜在的周期模式。后通过分析频谱图中的峰值来确定主要周期。最后绘制频谱图，标识出显著的周期频率。

Step3：季节性分析：首先分析月平均水沙通量，观察其季节性变化，后计算季节因子理解不同月份之间差异，最后绘制折线图展示季节性变化趋势。

Step4：突变型分析：通过 MK 检验检测时间序列中是否存在趋势。输出最后检验的结果所给出的 p 值，通过其与预先设定的显著性水平 0.05 比较，确定其是否有趋势，最后输出 $u_f(k)$ 和 $u_b(k)$ 图形更清晰的确定确定突变所在位置，分析突变规律。

6.3 求解结果

6.3.1 数据处理

通过数据缺失值处理后，对其提取平均值、标准差、最大值、最小值、峰值和偏度六个数据特质，并将结果输出于下表：

	平均值	标准差	最小值	最大值	偏度	峰度
水通量	27.596561	21.12779273	5.202576	108.24768	1.458	2.244
沙通量	16.93337852	28.17617637	0.24880459	135.0561773	2.563	6.657

6.3.2 周期性分析

依据问题求解，首先对水通量数据进行傅里叶变换，输出如下频谱图：

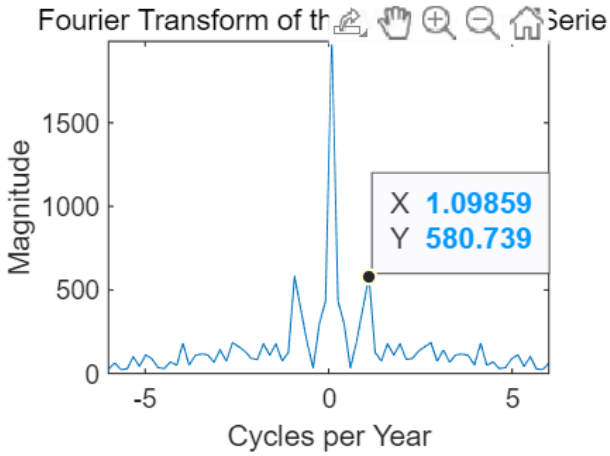

图 5　未去均值水通量频谱图

分析可得由于信号中的直流分量干扰，导致频谱图中表现一个非常大的峰值，掩盖了其他频率分量，导致周期性难以观察，所以决定对频谱图进行去均值处理，输出结果如下图：

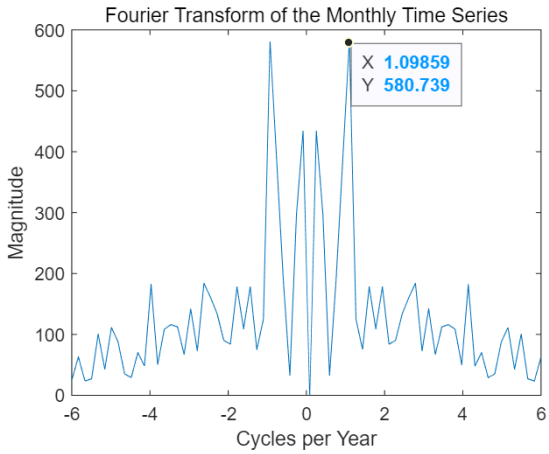

图 6　去均值后水通量频谱图

分析可得具有明显周期性，周期为一年，模型建立成功对沙通量数据进行同样操作输出如下频谱图：

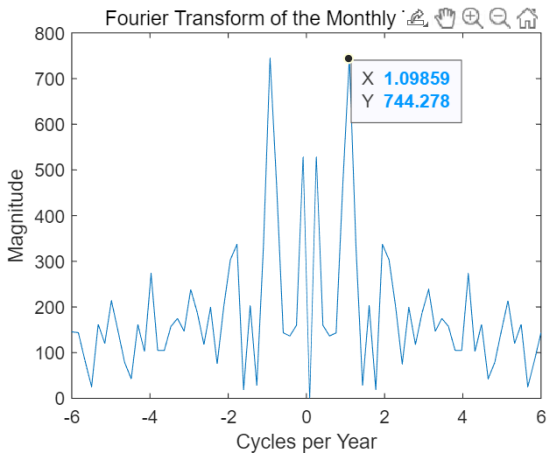

图 7　去均值后沙通量频谱图

分析可得具有明显周期性，周期为一年，综上水文站水沙通量的周期性如下表所示：

表 7　水文站水沙通量周期性规律

数据	周期
水通量	一年
沙通量	一年

6.3.3 季节性分析

依据问题求解首先计算月平均水沙通量，并将其可视化如下：

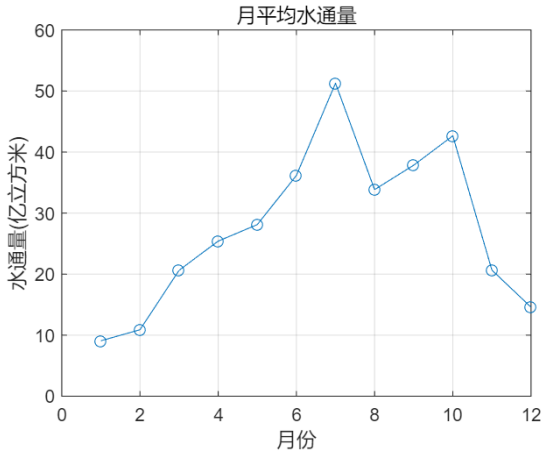

图 8　月平均水通量

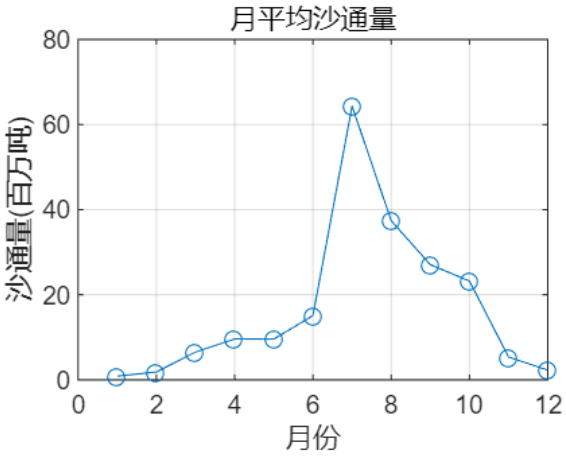

图 9　月平均沙通量

分析可得其存在较为明显的季节性，求解其季节因子以突出季节性，结果如下表:

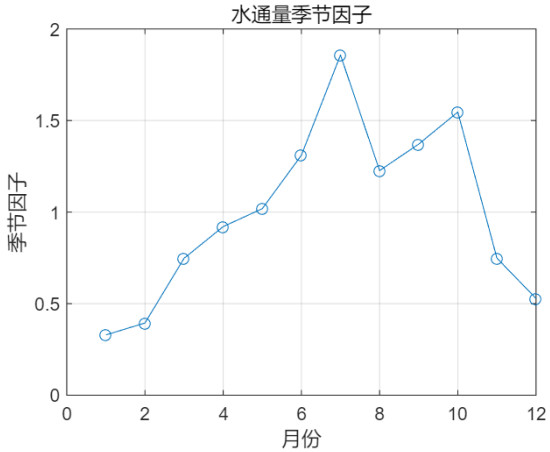

图 10　水通量季节因子

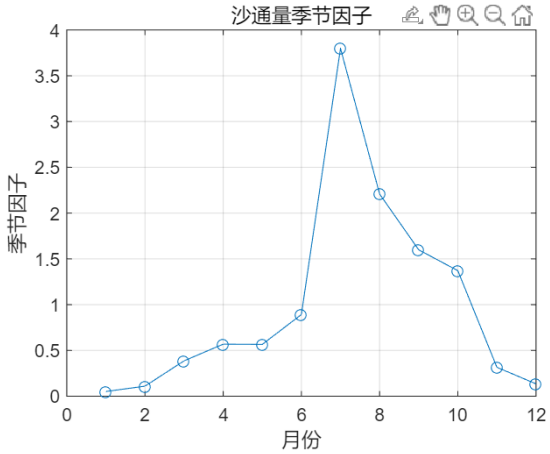

图 11　沙通量季节因子

通过观察图像可得，水通量在 6-10 月份较多，在 12-2 月份较少，沙通量在 7-10 月份较多，11-3 月较少，通过以上图像我们可以明显分清其季节性。

6.3.4 突变型分析

依据问题求解首先计算月平均水沙通量，后通过月平均水沙通量进行 MK 检验输出 p 值如下表所示：

表 8　水沙月通量 M-K 检测结果

水月通量 M-K 检验结果	p=1.68176e-06
沙月通量 M-K 检验结果	p=5.12246e-07

分析可得水沙通量存在突变型，输出 $u_f(k)$ 和 $u_b(k)$ 图形以确定突变点：

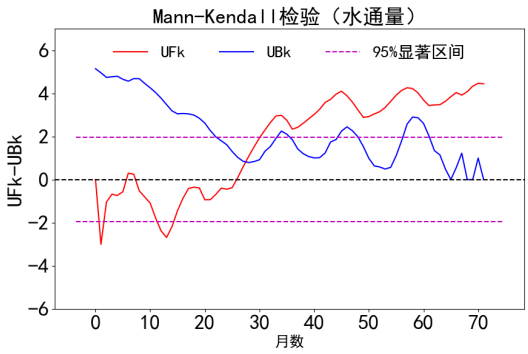

图 12　水月通量 M-K 检验

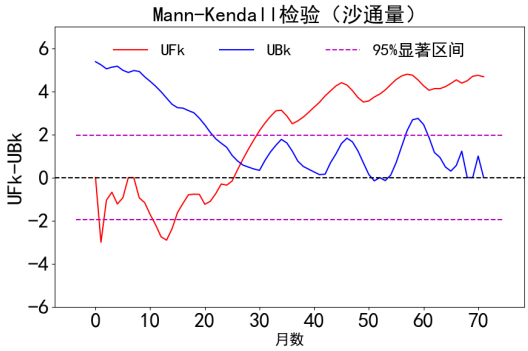

图 13　沙月通量 M-K 检验

分析可得水文站水沙通量数据具有突变点，其结果如下表：

表 9　水文站水沙通量突变规律

数据	突变所在月份	突变所在时间
水通量	28	18 年 4 月
沙通量	27	18 年 5 月

七、 问题三的模型的建立和求解

7.1 模型建立

7.1.1 R/S 分析方法

R/S 分析方法是一种用于时间序列分析的技术，它主要用于检测和量化时间序列数据中的长期依赖性或自相似性。通过其可以检验本组时间序列模型是否未来存在相同趋势、反转趋势或随机趋势。具体实现如下：

1. **构建累积总和序列：**首先，给定一个时间序列 $\{X_t\}_{t=1}^{N}$，计算该序列的平均值 μ。然后，构建一个新的时间序列 $\{Y_t\}_{t=1}^{N}$ 定义为：

$$Y_t = \sum_{i=1}^{t}(X_i - \mu) \tag{5}$$

   这称为累积偏差序列。

2. **计算 R/S 统计量：**对于累积偏差序列 $\{Y_t\}_{t=1}^{N}$，计算每个子区间上的范围 R，即最大值与最小值之差。计算每个子区间上的标准差 S，即原始序列 $\{X_t\}_{t=1}^{N}$ 的标准差。定义 R/S 统计量为：

$$\frac{R}{S} \tag{6}$$

3. **重复计算不同长度的子区间：**将原始序列分成不同长度的子区间，并对每个子区间重复上述步骤，从而得到一系列 R/S 值。
4. **绘制 R/S 统计量的对数-对数图：**绘制子区间长度 n 的对数与相应的 R/S 统计量的对数之间的关系图。如果时间序列表现出自相似性，则该图应呈现出一条直线。
5. **计算 Hurst 指数：**

   利用线性回归来拟合对数-对数图，斜率即为 Hurst 指数 H。Hurst 指数的取值范围是 0 到 1，不同的值代表了不同的行为特征：如果 $H < 0.5$ 表示反持续性（anti-persistent），即序列倾向于反转；如果 $H = 0.5$ 表示随机行走（random walk），即序列是随机的；如果 $H > 0.5$ 表示持续性（persistent），即序列倾向于保持趋势。

7.1.2 SARIMA 模型

SARIMA 模型是一种扩展的 ARIMA 模型，ARIMA 模型由三个部分组成：自回归（AR）、差分（I）、移动平均（MA）。SARIMA 模型除了这三个非季节性组件之外，还包含季节性的 AR、I、MA 成分。包含了额外的季节性成分。因此其专门用来处理具有季节性模式的时间序列数据。针对水沙通量的预测，考虑到时间序列具有季节性等特征，所以决定应用该模型。

SARIMA 模型通常表示为 $SARIMA(p,d,q)(P,D,Q)_m$ 其中：p: 非季节性自回归项的阶数。d: 非季节性差分的阶数。q: 非季节性移动平均项的阶数。P: 季节性自回归项的阶数。D: 季节性差分的阶数。Q: 季节性移动平均项的阶数。m: 季节性的周期。具体实现如下：

1. 数据可视化：分析观察时间序列数据，了解其特性，如趋势、季节性等。
2. 分解时间序列：使用季节性分解等方法来分离出趋势和季节性成分。
3. 确定非季节性部分：使用自相关函数（ACF）和偏自相关函数（PACF）来确定非季节性 ARIMA 模型的参数 p, d, q。
4. 确定季节性部分：对季节性数据进行同样的分析来确定季节性部分的参数 P, D, Q。
5. 模型拟合：使用确定的参数来拟合 SARIMA 模型。
6. 诊断模型：时间序列预测之后，应对结果进行残差分析，目的首先是检查模型是否正确地捕捉到了时间序列的主要特征、验证模型假设是否合理。然后可以量化模型预测的准确性，了解模型预测值与实际观测值之间的偏差程度。再者可以发现模型中的潜在问题，例如非线性关系、异方差性或自相关等。同时通过残差图可以识别模型可能未捕捉到的模式。最后残差分析通常涉及对残差进行正态性检验，可以确保残差满足回归分析的基本假设，比如残差应该是独立同分布的（IID），并且最好服从正态分布。对 SARIMA 模型预测数据进行残差分析具体实现如下：

   (a) 残差计算

   假设我们有一个时间序列 $\{Y_t\}_{t=1}^{T}$，并且使用 SARIMA 模型对这个序列进行了拟合和预测。设 $\hat{Y}_t$ 是在时刻 t 的预测值，那么残差 e_t 可以表示为：

$$e_t = Y_t - \hat{Y}_t$$

   (b) 计算残差统计性质

   - **均值**：残差的均值应该接近于零。

$$\bar{e} = \frac{1}{T}\sum_{t=1}^{T} e_t \approx 0$$

- **方差**：残差的方差应该稳定，无显著变化。

$$\sigma_e^2 = \frac{1}{T-1}\sum_{t=1}^{T}(e_t - \bar{e})^2$$

- **正态性**：残差应该大致服从正态分布。

$$e_t \sim \mathcal{N}(0, \sigma_e^2)$$

(c) 进行自相关检验

- **自相关函数（ACF）**：计算残差的一阶自相关系数 r_1 和更高阶的自相关系数。

$$r_k = \frac{\sum_{t=k+1}^{T}(e_t - \bar{e})(e_{t-k} - \bar{e})}{\sum_{t=1}^{T}(e_t - \bar{e})^2}$$

- **偏自相关函数（PACF）**：计算残差的偏自相关系数。

(d) 进行偏自相关检验

- 定义多元线性回归模型对于给定的滞后 k，我们定义多元线性回归模型如下：

$$Y_t = \beta_0 + \sum_{j=1}^{k}\beta_j Y_{t-j} + \varepsilon_t,$$

其中 β_0 是截距项，β_j 是对应于第 j 个滞后的系数，ε_t 是误差项。

- 估计回归系数通过最小二乘法估计上述模型中的参数 $\beta_0, \beta_1, \ldots, \beta_k$。
- 计算偏自相关系数偏自相关系数 ϕ_{kk} 被定义为上面模型中对应于第 k 个滞后项的系数 β_k，即：

$$\phi_{kk} = \beta_k.$$

偏自相关系数的计算涉及到多元回归的估计，可以使用如下的矩阵形式表示：

$$\phi_{kk} = \left(\mathbf{R}_k^{-1}\right)_{kk},$$

其中 $\mathbf{R}_k$ 是由自相关系数构成的 $k \times k$ 的 Toeplitz 矩阵，$\left(\mathbf{R}_k^{-1}\right)_{kk}$ 表示该逆矩阵的第 k, k 个元素。

- 重复步骤 1 至 3 对于每个所需的滞后 k，重复以上步骤以获得对应的偏自相关系数。

7. 模型选择：根据残差分析的结果调整模型参数或选择更合适的模型形式，改善模型的预测能力，同时使用信息准则（例如 AIC 或 BIC）来比较不同的模型，并选择最优模型。

7.1.3 小波分析

小波分析是一种时间-频率分析方法，它可以同时提供信号在时间和频率两个维度上的信息。与传统的傅里叶变换不同，小波分析可以在不同的尺度上捕捉信号的局部特征，不仅方便实现分析多尺度特性的信号，又可以提供信号在时间上的局部细节，帮助发现特定时间段内的异常变化。因此决定最终对 SARIMA 预测模型生成数据应用小波分析法，分析旱涝灾害时期，并合理分配监测次数。

小波分析的核心思想是使用可伸缩的小波基函数来表示信号的不同部分。小波函数通常具有局部支持，并且可以缩放和平移以适应信号的不同特征，帮助提取信号在不同尺度上的具备特征，分析预测模型生成数据决定应用离散小波变换（DWT) 其主要通过递归地将信号分解为近似系数和细节系数来实现，对于一维信号，DWT 可以表示为：

$$c_j(k) = \sum_{n=-\infty}^{\infty} f(n)h(n-2k) \tag{7}$$

$$d_j(k) = \sum_{n=-\infty}^{\infty} f(n)g(n-2k) \tag{8}$$

其中：

- $c_j(k)$ 是第 j 层的近似系数。
- $d_j(k)$ 是第 j 层的细节系数。
- $f(n)$ 是输入信号。
- $h(n)$ 和 $g(n)$ 分别是低通滤波器和高通滤波器的系数。
- j 是分解级别，k 是时间索引。

小波函数是小波分析中的核心组成部分，常见的小波函数包括 Haar 小波、Daubechies 小波、Meyer 小波等。这里以 Daubechies 小波为例，对其尺度函数 $\phi(x)$ 和小波函数 $\psi(x)$ 进行分析：对于 Daubechies 小波 db2（即 $N = 2$），尺度函数和小波函数可以表示为：

$$\phi(x) = \sum_{n=-\infty}^{\infty} h(n)\phi(2x-n) \tag{9}$$

$$\psi(x) = \sum_{n=-\infty}^{\infty} g(n)\phi(2x-n) \tag{10}$$

其中 $h(n)$ 和 $g(n)$ 分别是低通滤波器和高通滤波器的系数，对于 db2 小波，这些系数可以通过预测数据计算。

7.2 模型求解

通过对问题三进行分析确定模型通过 R/S 分析、SARIMA 预测、小波分析三部分完成，具体实现如下：

Step1： R/S 分析：对数据进行 R/S 分析，计算 R/S 统计量与 Hurst 指数，对其分析确定水文站水沙通量变换趋势。

Step2： SARIMA 预测：采用 SARIMA 预测模型，以前五年数据为训练集，2021 年数据为验证集，对 2022-2023 年数据进行预测分析，并对最终结果可视化，最后通过残差分析评估模型性能验证模型假设的有效性，分析并验证预测结果。

Step3： 小波分析：应用小波分析法处理 SARIMA 预测数据，分析旱涝灾害所在时期，并制定最优采样检测方案确定采样检测次数与具体时间。

7.3 求解结果

7.3.1 R/S 分析

依据问题求解对附件提供数据进行 R/S 分析，计算 R/S 统计量与 Hurst 指数结果如下表：

数据	Hurst 指数	R/S 值
水月通量	0.83857	19.34843
沙月通量	0.76523	15.87536

分析可得 $H > 0.5$ 表示时间序列具有持续性（persistent），倾向于保持趋势。

7.3.2 SARIMA 预测

依据模型求解对数据进行预测分析，预测结果见附件 2，预测结果见下图：

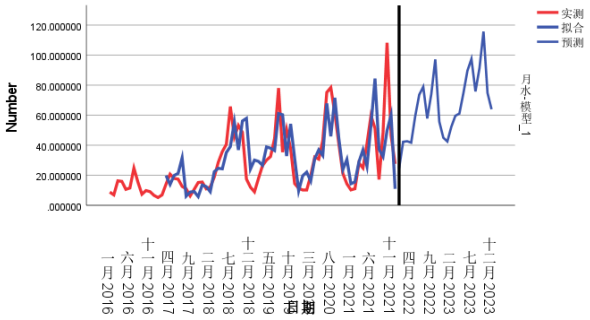

图 14　水通量实测与预测对照图

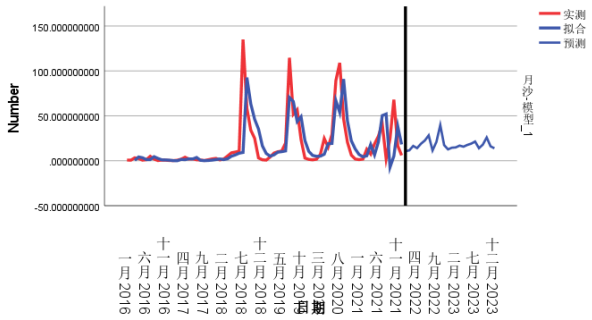

图 15　沙通量实测与预测对照图

分析可得水沙通量预测数据与实测数据重合性较强，说明模型建立较为成功，因此构建残差图进一步对模型进行评估：

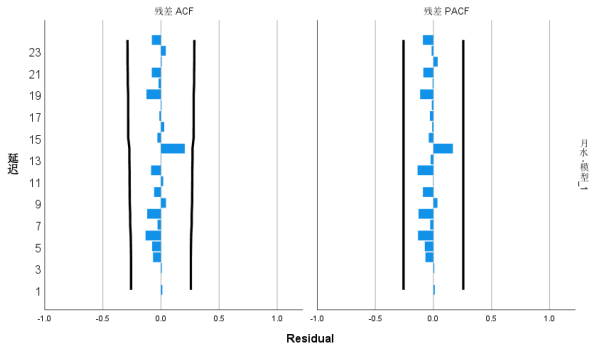

图 16　沙通量残差分析

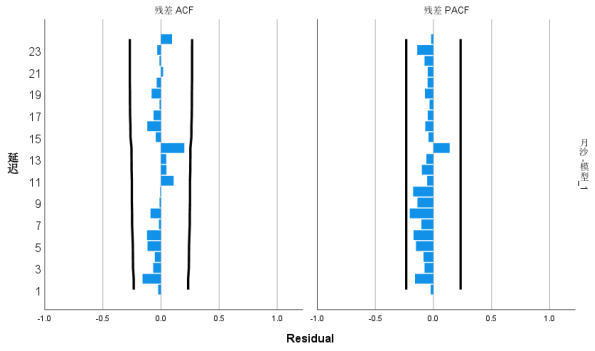

图 17　沙通量残差分析

下面对残差图进行分析：

1. 水通量残差分析

(a) 自相关函数图

分析可得残差自相关性并不明显，大部分点位于置信区间以内，表明残差之间没有明显的线性关系。

(b) 偏自相关函数图

分析除了滞后 1 个单位外，其他点都落在置信区间内。这意味着残差之间的关联主要体现在滞后一个单位上，而在更长的时间间隔上则不明显。这种现象可能是由于模型中的某些因素未被充分捕捉导致的，例如季节效应或其他周期性影响。

综上模型假设基本成立。但考虑到 ACF 图中滞后 1 到 4 个单位的异常点，可能存在一些未被模型捕获的因素。

2. 沙通量残差分析

(a) 自相关函数图分析得其前几个滞后期呈现负相关性，并且随着滞后期数增加逐渐减小到接近零。这表明可能存在一定的短期依赖性。然而，在较长的滞后期内，残差自相关函数值变得非常小或接近于零，说明其长期依赖性较弱。

(b) 偏自相关函数图残差偏自相关函数图显示在第一个滞后期后，残差偏自相关函数值迅速下降并趋于平稳。这意味着尽管在第一个滞后阶数处残差之间存在较强的关联性，但这种关联性很快就会消失，即后续的滞后项对当前时刻的残差没有显著的影响

综上模型假设基本成立，模型短期依赖性强，长期依赖性较弱。

7.3.3 小波分析

根据问题求解对 SARIMA 预测数据进行小波分析，输出结果如下图：

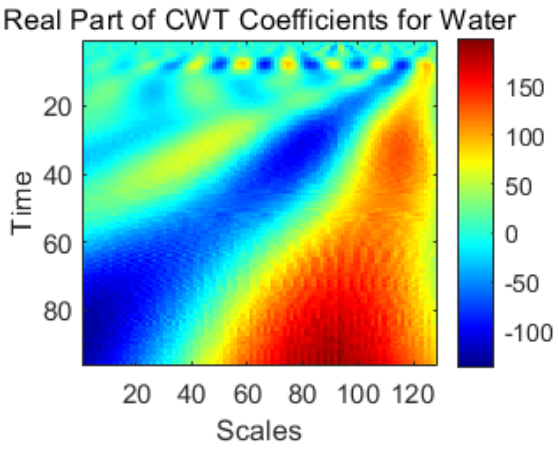

图 18　水通量小波系数等值线图

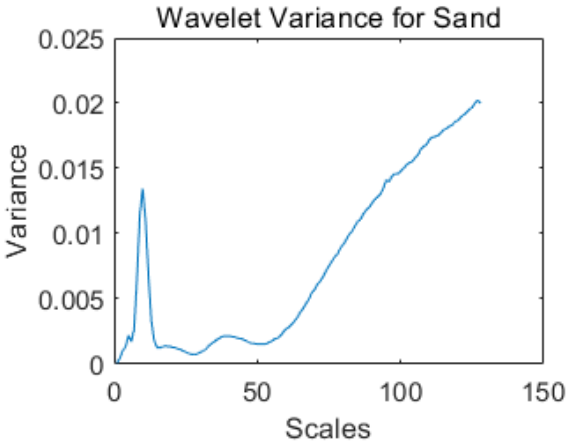

图 21　沙通量小波方差图

l

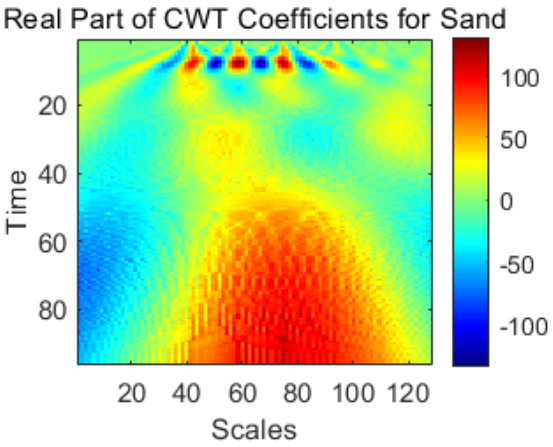

图 19　沙通量小波系数等值线图

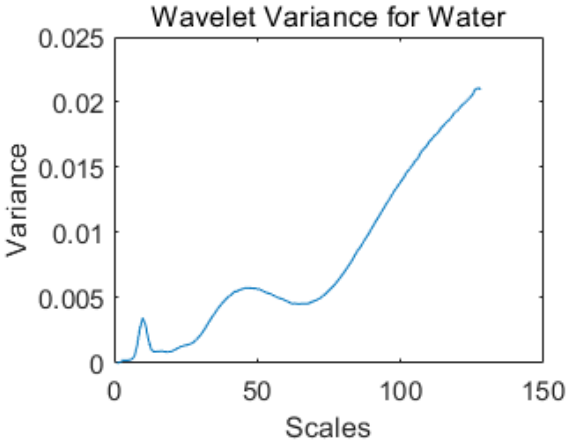

图 20　水通量小波方差图

可以发现秋季时段，水沙通量均具有一定的稳定性，所以可以减少测量次数，例如

3 天一次，夏天 5 月份左右，一天一次，6-7 月一天两次春冬季节，采取 2 天一次的测量方式。

八、 问题四的模型的建立和求解

8.1 模型建立

8.1.1 河底高程分析

需要分析河底高程来评价调水调沙的实际效果，选取了河底高程均值为某一测量时间的特征数据，且由于每次测量的范围不同，我们选取了一段共有的区间作为研究对象。并依据以下过程进行平均高程求解。

第 1 步： 把数据点 (x_i, y_i) 按照起点距x 从小到大排序，$i = 1, 2, ..., n.$

第 2 步： 计算相邻两个起点距的长度： $\Delta x_i = x_i - x_{i-1}, \quad i = 2, 3, ..., n.$

第 3 步： 计算相邻两个河底高程的平均高度： $\overline{y}_i = 0.5(y_i + y_{i-1}), \quad i = 2, 3, ..., n.$

第 4 步： 计算面积微元： $dS = \Delta x_i \overline{y}_i, \quad i = 2, 3, ..., n.$

第 5 步： 计算总面积： $S = \sum_{i=2}^{n} \Delta x_i \overline{y}_i.$

第 6 步： 计算平均河底高程： $h = S/a$.

其中 a 为起点桩与终点桩之间的距离。根据附件 2 数据，得到测量时间的平均高程。计算结果如下表所示：

日期	2016-6-8	2016-10-20	2017-5-11	2017-9-5	2018-9-13	2019-4-13	2019-10-15	2020-3-19	2021-3-14
河底平均高程	45.0753	45.1617	45.2457	45.3311	45.3401	45.0251	44.8647	44.9585	44.9832

对其数据进行可视化得到如下图：

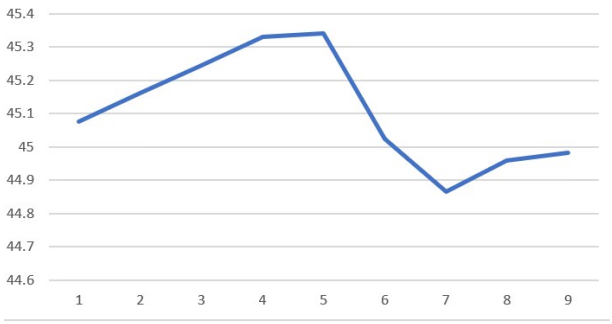

图 22　平均高程可视图

通过观察上表和上图，不难发现，在未进行“调水调沙”下，河底高程日益上升，这有害于河两岸的居民的生存环境。在进行“调水调沙”后，每年的高程维持在 45m 左右，表明了“调水调沙”对于河底高程的效果。

8.1.2 水沙通量分析

需要分析水沙通量来评价“调水调沙”作用时，我们首先可以从分析其数据入手，下图为水通量和沙通量的示意图。

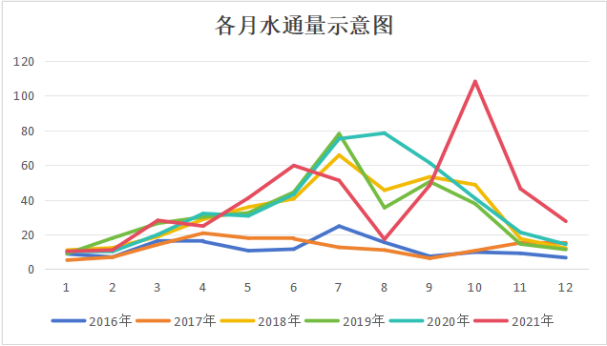

图 23　各月水通量示意图

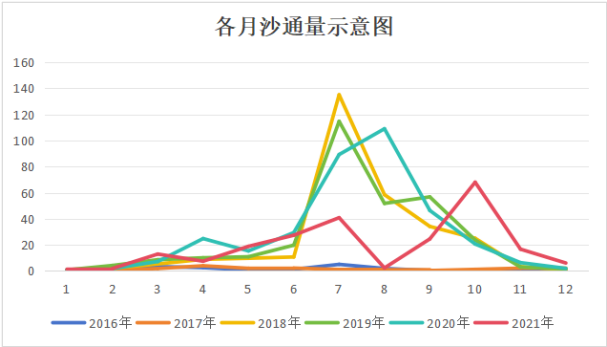

图 24　各月沙通量示意图

综合这两张图，可以看出，各月的水通量于 6-7 月均较大，一般而言，较大的水通量应能带走更多的沙子，减少河底高程进一步增加，然而观察 2016、2017 年的沙通量在 6-7 月份并没有于其他月份有太大区别，这解释了为什么上文中未进行“调水调沙”下，河底高程会增加。

分析 2018 年以后的 6-7 月份数据，我们不难发现，沙通量发生了明显的改善，这表明河底中更多沉沙被带入海中，所以其河底高程发生下降，其沙通量变化更符合水通量变化，很好说明了“调水调沙”对于水沙通量的作用。

8.1.3 未“调水调沙”的河底高程预测

要预测未经过“调水调沙”的河底高程，我们需要先分析 2016-2018.5 月之前的高程数据，以来分析之后的数据，通过观察图 20 不难发现，其未经过“调水调沙”的河底高程基本呈线性关系增加，所以我们通过前四份时间数据，利用最小二乘法拟合得到其关系后预测 2026 年高程，其结果为 46.30m。

下图为预测结果可视化：

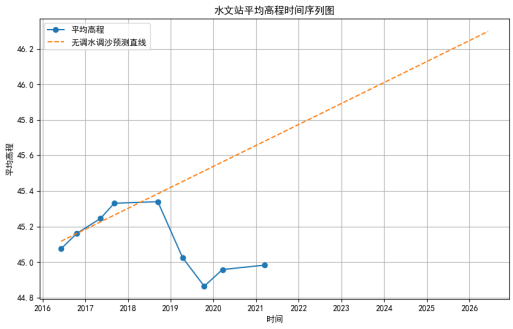

图 25　预测结果

8.1.4 评价“调水调沙”的重要性

“调水调沙”作为一种重要的水利工程措施，主要应用于治理水沙关系不协调的河流，尤其是像黄河这样的河流，黄河由于特殊的地理环境和气候条件，形成了水少沙多的特点，这导致了水沙关系的不协调，严重影响了河流的健康和稳定，“调水调沙”可以帮助黄河缓解这种不协调问题。同时其还可以防止河床抬升、保障防洪安全、维持生态稳定和提高水资源利用效率。

如果不进行“调水调沙”，10 年后的河底将比“调水调沙”稳定的 45m 高出 1 米多，其势必会造成大范围的灾祸，所以“调水调沙”是一项伟大的工程。

九、 模型的评价

9.1 模型的优点

- 优点 1 灵活性高：应用模型多允许根据实际情况灵活选择调整参数，以适应各种类型的数据。
- 优点 2 精准度高：模型建立过程中均引入相关参数进行评价，并根据调整不断优化，保证参数保持最佳。
- 优点 3 引用范围广可直接应用到其他类似模型封装性较好。

9.2 模型的缺点

- 缺点 1 参数选择困难，提供参数选择可能性多，参数选择较为复杂，对专业知识和经验要求较高。
- 缺点 2 数据要求较高，应用拟合算法与 SARIMA 模型等多种算法前者依赖数据迭代以降低误差，后者依赖大量数据集进行训练。

参考文献

[1] 司守奎, 孙玺菁. 数学建模算法与应用[M]. 北京: 国防工业出版社, 2011.

[2] 卓金武. MATLAB 在数学建模中的应用[M]. 北京: 北京航空航天大学出版社, 2011.

[3] 朱厚华, 秦大庸, 周祖昊. 黄河流域降雨量时间序列演变规律分析[C]//中国自然资源学会 2004 年学术年会论文集 (上册). 南京: 中国水利水电科学研究院水资源研究所(北京), 2004: 516-523.

[4] 郎国放. 山东省月度财政收入的时间序列分析[J]. 山东经济, 2005, 21(6):45-49.

[5] 孙忠保. 近 60 年淮河流域极端降水和极端温度时空变化特征[J]. 贵州师范大学学报（自然科学版）, 2024, 42(3):35-45.

[6] 郭彦, 侯素珍, 王平, 等. 基于小波分析的黄河上游水沙多时间尺度特征[J]. 干旱区研究, 2015, 32(6):1047-1054.

[7] “黄河流域水系统治理战略与措施”项目组. 黄河流域水系统治理战略研究[J]. 中国水利, 2021(5):1-4.

[8] 任剑飞. 基于流速面积法的河流水位流量计量研究[J]. 水利规划与设计, 2015(11): 69-71.

[9] 黄昊, 邓应梅. 基于 sARIMA 模型的某三甲医院神经外科手术工作量预测分析[J]. 中国病案, 2023, 24(11):52-54.

[10] 方惟一, 郝文渊, 祖力皮卡尔 · 吐迪, 等. 2011-2020 年乌鲁木齐市手足口病流行病学特征分析及 SARIMA 预测模型构建[J]. 中国医药导报, 2024, 21(16):24-28.

附录 A　文件列表

文件名	功能描述
main2.py	问题二程序代码
main3.py	问题三程序代码
main4.py	问题四程序代码
pinpu.m	绘制频谱图程序代码
wavelet.m	绘制小波等值线程序代码

附录 B　代码

main2.py

```python
from matplotlib.ticker import FuncFormatter
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import rcParams
import matplotlib.dates as mdates
rcParams['font.sans-serif'] = ['SimHei']
rcParams['axes.unicode_minus'] = False


file_path = 'pb2\六年水沙日通量.csv'
df = pd.read_csv(file_path, index_col=0)
df.columns = ['Year', 'Month', 'Day', 'WaterFlux', 'SandFlux']

df['WaterFlux'] = df['WaterFlux'] / 100000000  # 亿立方米
df['SandFlux'] = df['SandFlux'] / 1000000000  # 百万吨

df['Datetime'] = pd.to_datetime(df[['Year', 'Month', 'Day']])
df['年份'] = df['Datetime'].dt.year
df['月份'] = df['Datetime'].dt.month

plt.figure(figsize=(10, 6))
```

```
plt.plot(df['Datetime'], df['WaterFlux'], label='水通量(亿立方米)')
plt.plot(df['Datetime'], df['SandFlux'], label='沙通量(百万吨)')
plt.xlabel('时间')
plt.ylabel('通量')
plt.title('水沙通量时间序列图')
plt.legend()
def format_date(x, pos=None):
    date = mdates.num2date(x)
    if date.month == 1:
        return date.strftime('%Y-%m')
    else:
        return date.strftime('%m')
plt.gca().xaxis.set_major_formatter(FuncFormatter(format_date))
plt.gca().xaxis.set_major_locator(mdates.MonthLocator())
plt.gcf().autofmt_xdate()
plt.xticks(fontsize=7)
plt.savefig('pb2\水沙通量时间序列图.svg')



plt.show()
```

main3.py

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import linregress
from matplotlib import rcParams
from tqdm import tqdm
from statsmodels.tsa.statespace.sarimax import SARIMAX
from sklearn.metrics import mean_absolute_error,
    mean_squared_error, r2_score
```

```
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
rcParams['font.sans-serif'] = ['SimHei']
rcParams['axes.unicode_minus'] = False

file_path = 'pb3\六年水沙月通量.csv'
monthly_flux = pd.read_csv(file_path, encoding='utf-8-sig')
monthly_flux['DATE'] = pd.to_datetime(monthly_flux[['Year', '
    Month']].assign(DAY=1))
monthly_flux = monthly_flux.set_index('DATE')
monthly_flux = monthly_flux.drop(columns=['Year', 'Month'])
monthly_flux = monthly_flux[['水月通量', '沙月通量']]['2018-01
    ':'2021-12']

water_flux = monthly_flux['水月通量']
sand_flux = monthly_flux['沙月通量']

water_flux_diff1 = water_flux.diff().dropna()
sand_flux_diff1 = sand_flux.diff().dropna()
water_flux_diff2 = water_flux_diff1.diff().dropna()
sand_flux_diff2 = sand_flux_diff1.diff().dropna()

plt.figure(figsize=(14, 6))
ax1 = plt.subplot(4, 2, 1)
water_flux.plot(ax=ax1, label='水月通量原值')
plt.legend()
ax3 = plt.subplot(4, 2, 3)
water_flux_diff2.plot(ax=ax3, label='水月通量二阶差分')
plt.legend()
ax5 = plt.subplot(4, 2, 5)
plot_acf(water_flux_diff2, ax=ax5, lags=len(water_flux_diff2)
    /2 - 1)
ax7 = plt.subplot(4, 2, 7)
plot_pacf(water_flux_diff2, ax=ax7, lags=len(water_flux_diff2)
    /2 - 1)

```

```
ax2 = plt.subplot(4, 2, 2)
sand_flux.plot(ax=ax2, label='沙月通量原值')
plt.legend()
ax4 = plt.subplot(4, 2, 4)
sand_flux_diff1.plot(ax=ax4, label='沙月通量一阶差分')
plt.legend()
ax6 = plt.subplot(4, 2, 6)
plot_acf(sand_flux_diff2, ax=ax6, lags=len(sand_flux_diff2)/2 - 1)
ax8 = plt.subplot(4, 2, 8)
plot_pacf(sand_flux_diff2, ax=ax8, lags=len(sand_flux_diff2)/2 - 1)


# 拟合 SARIMA 模型
def fit_sarima(series, order, seasonal_order):
    model = SARIMAX(series, order=order, seasonal_order=seasonal_order)
    result = model.fit(disp=False)
    return result

# 水月通量的 SARIMA 模型
p, d, q = 4, 2, 1
P, D, Q, S = 0, 1, 1, 12
water_order = (p, d, q)
water_seasonal_order = (P, D, Q, S)
water_model = fit_sarima(water_flux, water_order, water_seasonal_order)

water_predict = water_model.get_prediction(start='2018-01', end='2023-12')
water_predict_mean = water_predict.predicted_mean
water_conf_int = water_predict.conf_int()
water_r2_score = r2_score(water_flux, water_predict_mean[:48])
print(f'{"="*10}水月通量预测{"="*10}')
```

```
print(water_predict_mean)
print(f'水月通量的 R² 值: {water_r2_score}')

plt.figure(figsize=(14, 6))
plt.plot(water_flux, label='水月通量原值')
plt.plot(water_predict_mean, label='水月通量预测值', linestyle='--')
plt.fill_between(water_predict_mean.index,
                 water_conf_int.iloc[:, 0],
                 water_conf_int.iloc[:, 1], color='k', alpha=0.1)
plt.legend()
plt.title('水月通量预测')

# 沙月通量的 SARIMA 模型
p, d, q = 1, 2, 2
P, D, Q, S = 1, 0, 0, 12
sand_order = (p, d, q)
sand_seasonal_order = (P, D, Q, S)
sand_model = fit_sarima(sand_flux, sand_order, sand_seasonal_order)

sand_predict = sand_model.get_prediction(start='2018-01', end='2023-12')
sand_predict_mean = sand_predict.predicted_mean
sand_conf_int = sand_predict.conf_int()
sand_r2_score = r2_score(sand_flux, sand_predict_mean[:48])
print(f'{"="*10}沙月通量预测{"="*10}')
print(sand_predict_mean)
print(f'沙月通量的 R² 值: {sand_r2_score}')

plt.figure(figsize=(14, 6))
plt.plot(sand_flux, label='沙月通量原值')
plt.plot(sand_predict_mean, label='沙月通量预测值', linestyle='--')
```

```
plt.fill_between(sand_predict_mean.index,
                 sand_conf_int.iloc[:, 0],
                 sand_conf_int.iloc[:, 1], color='k', alpha=0.1)
plt.legend()
plt.title('沙月通量预测')

predicted_result = pd.DataFrame({'水月通量': water_predict_mean[48:], '沙月通量': sand_predict_mean[48:]},
                                index=water_predict_mean[48:].index)
predicted_result.to_csv(r'pb3\2022-2023水沙月通量预测.csv', encoding='utf-8-sig')


plt.tight_layout()
plt.show()
```

main4.py

```
from matplotlib import rcParams
import pandas as pd
import matplotlib.pyplot as plt
rcParams['font.sans-serif'] = ['SimHei']
rcParams['axes.unicode_minus'] = False

file_path = 'pb4\部分监测点数据.csv'



scatter4 = data[['Date', 'Distance', 'Level', 'Depth']].copy()
scatter4 = scatter4.dropna(subset=['Distance', 'Depth'])
scatter4['Level'] = scatter4['Level'].ffill()
scatter4['Date'] = scatter4['Date'].ffill()
scatter4 = scatter4[scatter4['Depth'] != 0]
```

```
scatter4['Date'] = pd.to_datetime(scatter4['Date'])
scatter4['altitude'] = scatter4['Level'] - scatter4['Depth']

scatter4.to_csv('pb4\水站水底高程表.csv', index=False,
   encoding='utf-8-sig')

average_altitude = scatter4.groupby(scatter4['Date'].dt.date)[
   'altitude'].mean()

plt.figure(figsize=(10, 6))
average_altitude.plot(kind='line')
plt.title('Average Altitude over Time')
plt.xlabel('Date')
plt.ylabel('Average Altitude')
plt.grid(True)
plt.tight_layout()
plt.show()
```

pinpu.m

```
filePath = 'sj2.xlsx';

% 读取Excel文件
dataTable = readtable(filePath);



x = dataTable(:,2)

% 加载您的时间序列数据
N = height(x);
x = table2array(x)
% 对于按月采样的数据，我们假设每个月为一个采样点
% 因此，一年有12个月，即12个周期
% 这里我们定义一个"采样率"为每年12个样本
fs = 12; % 每年12个数据点
x_mean = mean(x);
```

```

x = x - x_mean;

% 应用FFT
X = fft(x);
X_mag = abs(X); % 幅度谱
X_mag = fftshift(X_mag); % 将零频移到中间

% 计算频率轴
f = linspace(-fs/2, fs/2, N);

% 绘制频谱图
figure;
plot(f, X_mag);
xlabel('Cycles per Year');
ylabel('Magnitude');
title('Fourier Transform of the Monthly Time Series');
```

wavelet.m

```
% 加载Excel文件
data = readtable('2016-2023.xlsx', 'Sheet', 1);

% 提取"水"和"沙"两列数据
water = data.water;
sand = data.sand;

% 执行连续小波变换 (CWT)
scales = 1:128; % 定义尺度范围
cwtCoefficientsWater = cwt(water, scales, 'morl'); % 使用
   Morlet小波
cwtCoefficientsSand = cwt(sand, scales, 'morl');

% 提取小波系数的实部
realCoefficientsWater = real(cwtCoefficientsWater);
realCoefficientsSand = real(cwtCoefficientsSand);
```

```
16
17 % 计算小波方差
18 waveletVarianceWater = sum(abs(realCoefficientsWater).^2, 2);
19 waveletVarianceWater = waveletVarianceWater / sum(
      waveletVarianceWater); % 归一化
20
21 waveletVarianceSand = sum(abs(realCoefficientsSand).^2, 2);
22 waveletVarianceSand = waveletVarianceSand / sum(
      waveletVarianceSand); % 归一化
23
24 % 调整数据顺序，使得纵轴从上到下递增
25 realCoefficientsWater = flipud(realCoefficientsWater);
26 realCoefficientsSand = flipud(realCoefficientsSand);
27
28 % 创建一个递减的时间标签向量
29 timeLabelsWater = length(water):-1:1;
30 timeLabelsSand = length(sand):-1:1;
31
32 % 绘制小波系数实部的等值线图
33 figure;
34 imagesc(scales, timeLabelsWater, realCoefficientsWater);
35 colormap(jet);
36 colorbar;
37 xlabel('Scales');
38 ylabel('Time');
39 title('Real Part of CWT Coefficients for Water');
40
41 figure;
42 imagesc(scales, timeLabelsSand, realCoefficientsSand);
43 colormap(jet);
44 colorbar;
45 xlabel('Scales');
46 ylabel('Time');
47 title('Real Part of CWT Coefficients for Sand');
48
```

```
% 绘制小波方差图
figure;
plot(scales, waveletVarianceWater);
xlabel('Scales');
ylabel('Variance');
title('Wavelet Variance for Water');

figure;
plot(scales, waveletVarianceSand);
xlabel('Scales');
ylabel('Variance');
title('Wavelet Variance for Sand');

% 显示图形
drawnow;
```