基于 SARIMA 时间序列模型预测黄河水沙数据变化

摘要

在黄河水沙通量变化模型构建中，变化趋势易受到季节变化所影响，对于季节性变化问题，如何选取合理差分是问题关键。本文研究使用Mann-Kendall方法检测黄河水沙通量突变性，利用SARIMA时间序列模型进行水沙数据预测，通过进行合理差分，调优模型性能，实现水沙通量变化预测。

对于问题一，通过分析z-score归一化后的时间折线统计图和二元数据线性折线图，找到变量之间的关系，得出三个量之间均有强相关性的结论。使用水流量，通过二元线性回归的方法，对含沙量进行按小时重采样，并对省略值进行回归预测，得到较为连续的数据，以精细数据的颗粒度。提出年排沙量，估算6年来每年的年总水流量和排沙量数据。

对于问题二，按每天水沙通量均值取样，使用Mann-Kendall检测突变天数，建立突变点的统计图。通过分析6年来连续的季节性水沙通量变化，列表得到不同季节的变化趋势，并分析每一段区间的增减变化规律，得到其周期性。

对于问题三，使用SARIMA模型进行预测，首先使用ACF对按季节取样的数据进行检测，使用解答二中的季节规律，对两种季节数据进行不同方式差分，得到较缓的数据进行不同年份的同月份直接差分，波动性较强的数据使用均值插补，计算同月份差分。进行ACF检测，得到p-value值接近于0时，将数据带入SARIMA模型中，搜索找到合理的参数，实现预测水沙变化。通过粗化采样颗粒，使用原始数据集对已有数据进行拟合，得到较为合理的监测方案。

对于问题四，删除原数据中6月和7月的值，并利用该年份6月份前到2016的数据，使用SARIMA预测未调水调沙的数据绘制占比图，比较二者数据差异，衡量效果。通过分析高程和未调水调沙的水沙通量关系，估计10年后的高程。

利用SARIMA模型可以很好的适应季节性带来的数据不稳定情况，通过利用回归对数据进行补值，丰富数据量，提高预测质量，利用Mann-Kendall检测突变天数，得到数据的突变周期，有利于预测。

关键词：SARIMA 时间序列预测 Mann-Kendall检测 水沙通量 季节性变化

一、问题重述

1.1 问题背景

黄河水沙通量的变化对黄河流域的环境治理、气候变化和人民生活有着巨大影响，通过研究黄河流域水沙通量来指导黄河流域水资源分配、协调人地关系、调水调沙、防洪减灾工作。

1.2 要解决的问题

（1）通过所给数据研究并建立该水文站黄河水的含沙量与时间、水位、水流量的关系，并估算 2016-2021 年该水文站的年总水流量和年总排沙量。

（2）分析近 6 年该水文站水流量和含沙量的突变性、季节性以及周期性之间的关系，得出水沙通量的变化规律。

（3）通过上述题目中研究的水沙通量变化规律，预测该水文站未来两年（2022、2023 年）水沙通量的变化趋势，为了能及时掌握水沙通量的变化情况，最大程度减少检测成本资源，制定未来两年最优的采样监测方案。

（4）通过分析该水文站每年 6-7 月的水沙通量和河底高程变化情况，评估小浪底水库在这个时间段内进行“调水调沙”对水文站的实际影响。如果小浪底水库不进行“调水调沙”，预测未来 10 年内该水文站的河底高程将如何变化。

二、问题分析

2.1 数据处理

分析附件一中数据可知，提供的含沙量数据有缺失值，为了方便后续分析含沙量与时间、水位、水流量之间的关系，使用均值插值法对数据进行数据清洗。利用清洗后数据进行多元线性回归，首先使用了z-score进行标准化数据，用来消除水流量和含沙量度量单位不同的区别，通过z-score之后可以使变量在后续模型中具有合理的权重。

2.2 问题一分析

题目一要求研究含沙量与时间、水位、水流量之间的关系，通过该关系估算出近六年来该水文站的年总水流量和年总排沙量。通过数据的预处理，使用连续的水流量对含沙量进行多元线性回归，以小时为单位进行采样，精细化采

样时间。通过得到的精细化含沙量数据从而对含沙量与时间、水位和水流量之间的三对关系进行分析。年水流量根据目前采样数据进行加和，年总排沙量与河流横截面积、流速以及含沙量成正相关关系。

2.3 问题二分析

由第一问中所估算的年总水流量和年总排沙量以及水流量和含沙量近六年变化趋势可以得出其周期性和季节性。利用了 Mann-Kendall 检测其突变性、变化规律。

2.4 问题三分析

题目中要求根据第二问水沙通量变化规律，分析未来两年的变化趋势。根据变化规律分析出 2016-2017 年水沙通量变化较小，2018-2021 年水沙通量变化较大，成时间变化趋势故使用 ARIMA 时间序列模型进行预测，然后在每一年中又有季节性变化，所以我们改用更加精确地 SARIMA 进行预测未来两年水沙通量的变化趋势。

2.5 问题四分析

题目要求根据水文站的水沙通量和河底高程的变化情况，分析每年 6-7 月小浪底水库进行“调水调沙”的实际效果。不进行“调水调沙”，10 年后该水文站的河底高程变化。水沙通量在问题二以及问题三中我们已经进行模型建立并分析求解。通过变化规律分析得出，每年 6-7 月的水流量、含沙量在一年中达到最大，通过小浪底水库的“调水调沙”，该时期的水沙通量、河底高程与平常时期差距减少。

三、模型假设

模型假设要点如下：

（1）假设水深和水流速均为 0 的点代表河流的两岸。

（2）假设河流水深为 40 米。

（3）假设水沙通量只考虑水流量和含沙量。

四、符号说明

序号	符号	含义
1	X	计算Z-score的数据向量
2	μ	数据集的平均值
3	σ	数据集标准差
4	Z	Z-score的结果值
5	β	多元线性回归参数
6	T	年总排沙量(t)
7	M	河流含沙量(kg/m^3)
8	S	横截面积(m^2)
9	V	水流流速(m/s)
10	Q	水流量(m^3/s)

五、模型建立与求解

5.1 问题一：分析含沙量与时间、水位和水流量之间关系，估算近六年来水文站的年总水流量和年总排沙量

5.1.1 数据预处理

首先，对附件一中的数据进行按日进行取样，通过分析流量与含沙量的关系，发现二者具有明显的正相关关系。将流水量数据按小时进行取样，对空缺值用两空缺值区间较大值填充，为了减少预测数据的重复值，提高数据质量，通过两实值之间的趋势，对填补的流水数据加减一定的微小值，且不影响数据整体结果。

利用得到的颗粒度更小的水流量数据，通过加减不影响结果的波动值，对含沙量使用多元线性回归预测，使用预测值填充按小时采样的含沙量的空缺值。

5.1.2 Z-score 标准化数据

对于流水量和测点含沙量同时以月进行采样，由于二者数值差距较大，为

方便在折线图像观测趋势和识别异常值，所以使用 Z-score 进行标准化。对于一系列流量和含沙量数据，二者的取样频率一致，月取样数 P 相同。因此二者 k 年中的数据向量 X 为：

$$X = \{x_1, x_2, x_3 \ldots x_n\}, n \in \{P * 12k\}$$

对于流量和含沙量数据而言，它们的 Z-score 值为：

$$\mathrm{Z} = \frac{X - \mu}{\sigma} \qquad (1-1)$$

每年度月份的 Z-score 流量和含沙量趋势图，如图 1 所示：

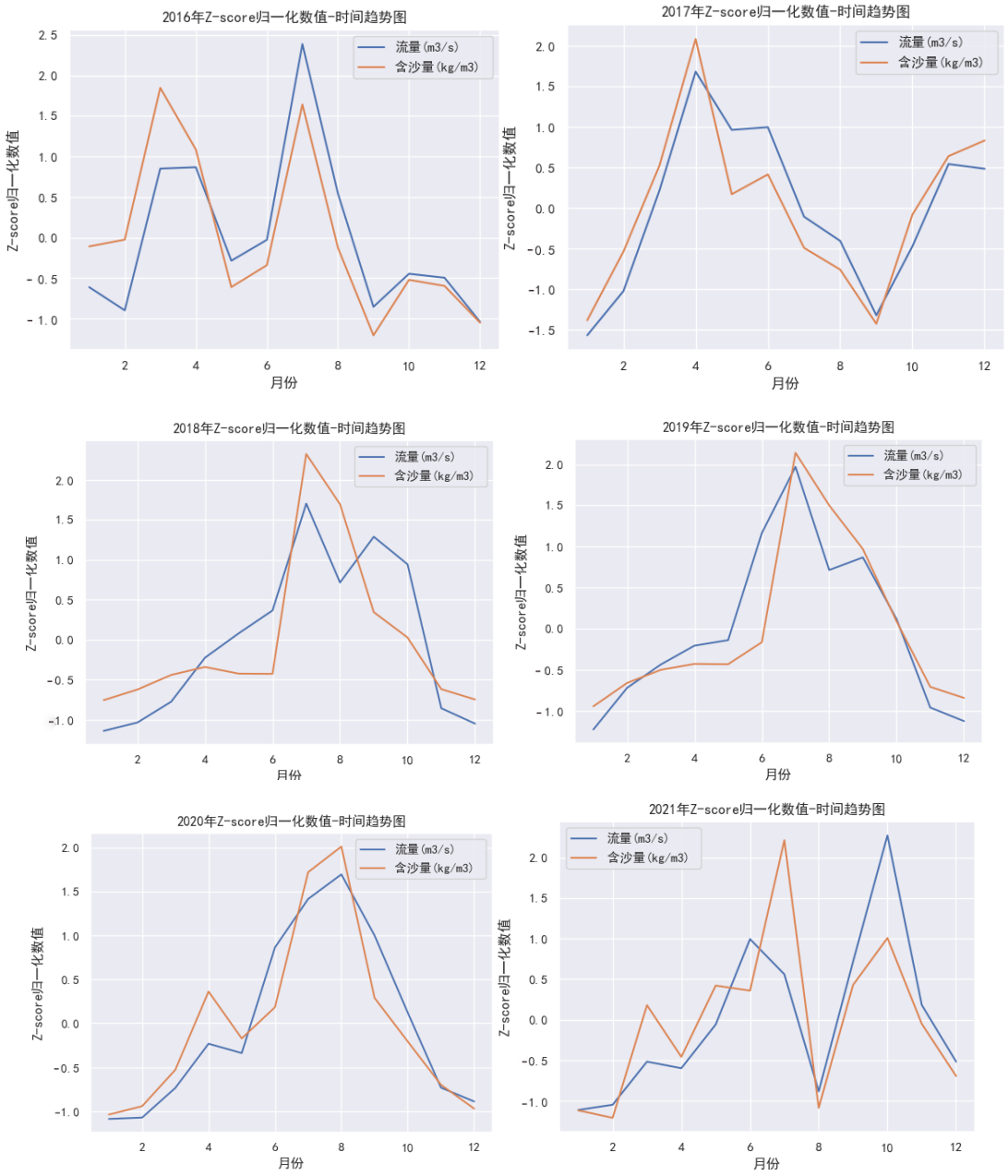

图1 2016-2021年流量、含沙量归一化后趋势图

估计流量和水位可能存在较强相关性，为方便衡量数据关系，因此对流量和水位进行图像曲线比较。水位趋势图和流量折线图如图 2 所示：

图 2 2016-2021 年流量和水位趋势图

因此判断二者有较强的相关性，在后续的含沙量测算中，使用流水量数据进行测算。

5.1.3 多元线性回归

设多元线性回归模型为 Y，截距为β_0，各自变量系数为$\beta_0, \beta_1, \beta_2..$，误差项为$\varepsilon$，即：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon \qquad (1-2)$$

绘制流量和含沙量的二元函数关系图，发现二者有明显的正相关性。如图 3 所示，泥沙量对水流量大致递增。

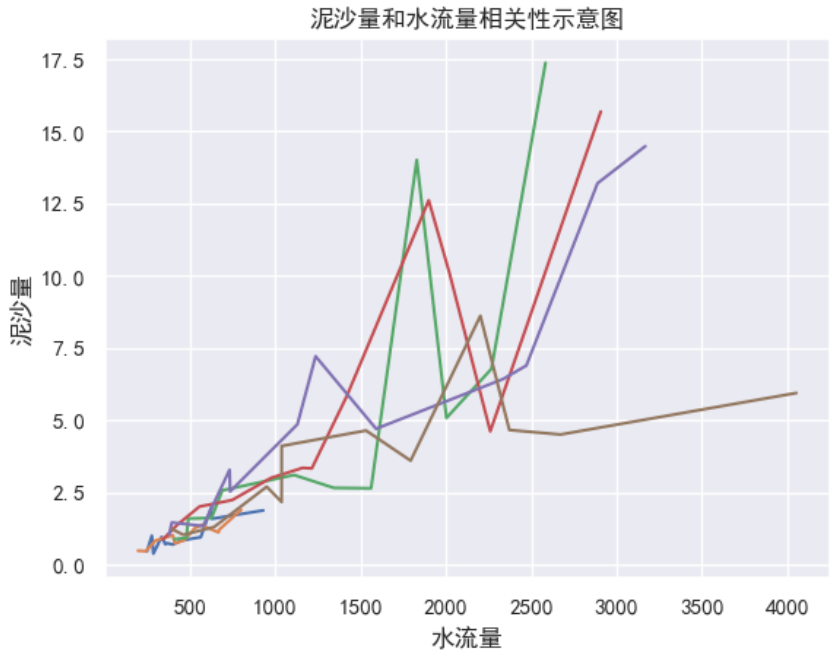

图 3 泥沙量和水流量相关性示意图

由于流速变化缓慢，为了方便后续模型建立，采取对流速取均值。根据附件 3，水深和水流速均为 0 的点代表河流的两岸。预计 1 月份和 7 月份河宽会有变化，分析两个月份的数据，差值不超过 100m，对整体流量数据影响较小，计算每两岸的两个干涸观测点得到河宽的平均直线距离为 378m，河流的横截面水高为水位减掉水深，水位水深变化不超过 20m，对数据影响不大，假设水高为 40 米。设 T 为排沙量，水流流速为 V，河流含沙量为 M，横截面积为 S，则：

$$T = V * M * S \qquad (1-3)$$

通过对小时取样的数据聚合到年取平均值，得到近2016-2021年的排沙量和水流量数据表，如表1和表2所示。

表1 2016-2021年的排沙量数据表

年份	2016 年	2017 年	2018 年	2019 年	2020 年	2021 年
排沙量（亿吨）	8.6	8.9	40.3	42.4	46.3	29.4

表2 2016-2021年水流量数据表

年份	2016 年	2017 年	2018 年	2019 年	2020 年	2021 年
水流量（亿立方米）	255	272	691	688	770	839

5.2问题二：分析水沙通量的突变性、季节性和周期性特征，研究其变化规律。

水沙通量为水流量和含沙量组成，根据已有数据建立2016年至2021年水流量和泥沙量的变化趋势图如图4所示，在图中我们可以清晰地得出2016-2017两年总量小，变化也小，从2018年开始，水流量与泥沙量都呈现明显的季节性、周期性变化趋势。

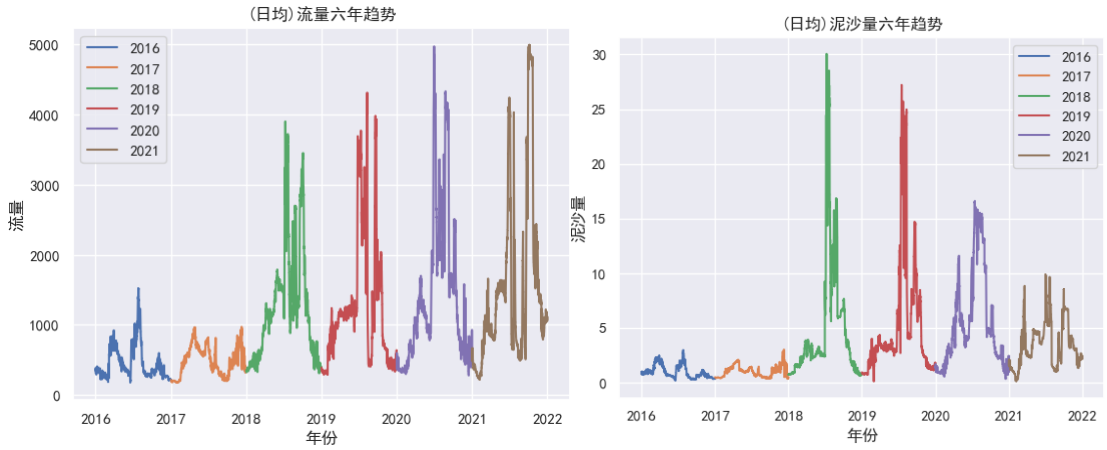

图4 近六年流量、含沙量变化趋势

通过分析上述图形数据得出在某一区间具有明显的递增性或递减性，且数据不遵循正态分布，所以我们引入常规的分析时间序列数据非参数检测方法，其对异常值不敏感，即使出现某些离群点也能得到较为可靠数据。

5.2.1 Mann-kendall突变性检测

对于河流水沙通量的突变性检测，采用$Mann-Kendall$方法进行统计说明，设置P-value值达到0.05时，两个变量具有变异的趋势。

$$Mann-Kendall统计量(S值)：S=\frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}sgn(x_j-x_i) \quad (1-4)$$

在上述公式中：

- n是时间序列数据的数据点数量。
- x_i和x_j是时间序列数据中的两个数据点，i和j分别表示数据点的位置。在此次模型检测中，我们采用了动态参数调参，将参数范围的左边界设置可滑动的概率递增，右边以整个数据集进行遍历递增。
- $sgn(x_j-x_i)$是一个符号函数，如果x_j大于x_i，则值为1。如果x_j小于x_i，则值为-1。如果x_j等于x_i，则值为0。

对于数据集中的所有可能的数据点对，计算它们差值的符号函数，将得出的函数值相加，并进行标准化，以获得一个度量趋势的统计量。如果S远离零，表明存在趋势，而越远离零，趋势越显著。

统计量S服从正态分布，其均值为0，方差 $Var(S)$ 按下式计算：

n≥8时，S大致服从正态分布:

$$Var(S) = \frac{n(n-1)(2n+5) - \sum_{i=1}^{k} t_i\,(t_i-1)(2t_i+5)}{18} \qquad (1-5)$$

通过Mann-kendall检测趋势得到突变性分析图如图5所示:

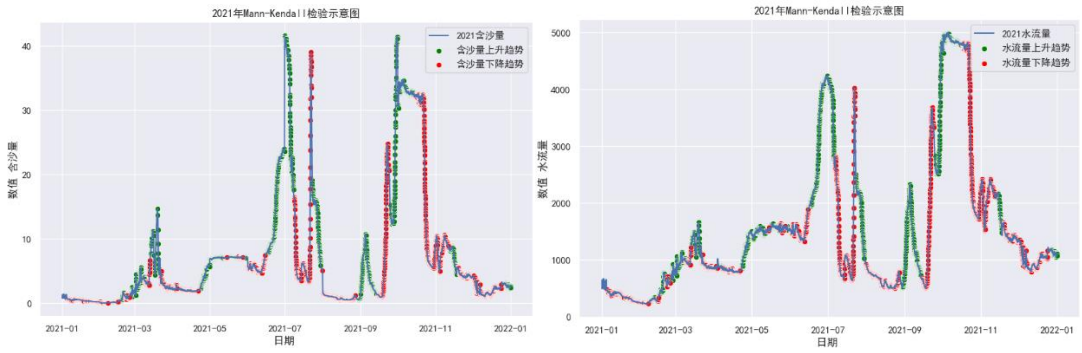

图5 2021年水流量和含沙量突变性分析

根据图4流量和含沙量变化规律以及图5突变性检测数据，总结出水沙通量变化规律，除2017年外，2016、2018-2021年份的4-7月份水沙通量总体上升，7-8月份所有年份水沙通量均呈下降趋势，整体而言，1-5月份上升趋势比较缓慢，5-7月份迅速上升，7-9月份整体下降，9-12月份较为平稳，10-11月份下降迅速，其余月份下降平缓，水沙通量的大体趋势如表3所示：

表3 2016-2021年水沙通量趋势表

<table>
<tr><th>月 \ 年</th><th>2016</th><th>2017</th><th>2018</th><th>2019</th><th>2020</th><th>2021</th></tr>
<tr><td>1-5 月份</td><td colspan="2">上升缓慢，4-5 月缓慢下降</td><td colspan="4">缓慢上升</td></tr>
<tr><td>5-7 月份</td><td colspan="2">上升缓慢，6-7 月有升有降</td><td colspan="4">迅速上升</td></tr>
<tr><td>7-8 月份</td><td colspan="2" rowspan="2">缓慢下降</td><td colspan="2">迅速下降</td><td>缓慢上升</td><td>迅速下降</td></tr>
<tr><td>8-9 月份</td><td colspan="2">缓慢上升</td><td>迅速下降</td><td rowspan="2">迅速上升</td></tr>
<tr><td>9-10 月份</td><td colspan="2">缓慢上升</td><td>缓慢下降</td><td colspan="2">迅速下降</td></tr>
</table>

10-11 月份		迅速下降
11-12 月份	平稳下降	迅速下降
12-次年 1 月	平稳回升	

5.3问题三：预测未来两年水沙通量的变化趋势，并制定最优的采样检测方案。

SARIMA具有季节性的自回归移动平均模型，简称季节性ARIMA。即在ARIMA的基础上，加入了季节性部分。季节性是指数据中具有固定频率的重复模式：每天、每两周、每四个月等重复的模式。

SARIMA模型可表示为SARIMA（p，d，q）x（P，D，Q）s，该式子满足乘法原则，前半部分表示非季节部分，后面表示季节部分，s表示季节性频率。

根据图4，得出数据具有季节性规律，同时其数据不平稳，使用SARIMA模型来捕捉季节性变化和趋势，同时使用表3的季节性规律将非平稳数据进行季节性差分，转为平稳数据。在差分阶数选择上，使用1阶差分，对于差分步数设置为每四个月一次，完成数据处理后，使用ACF对差分数据进行检测，检测结果如图6所示。

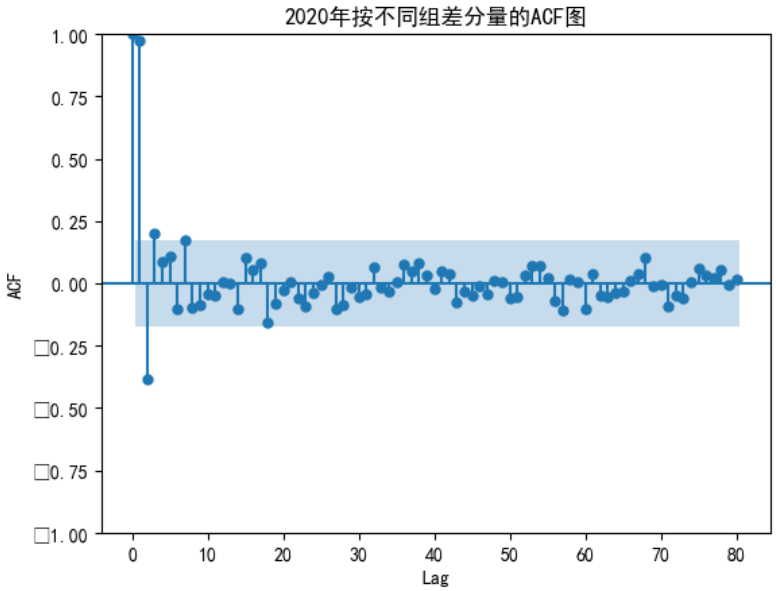

图6 2020年数据差分量的ACF图

图6中p值远远小于0.05，整体数据大体位于虚线以内，差分数据波动较弱。因而满足使用SARIMA模型进行预测的条件。

通过预测数据和原始数据的对比，如下图7 2022-2023未来两年预计和实际的水流量和泥沙量：

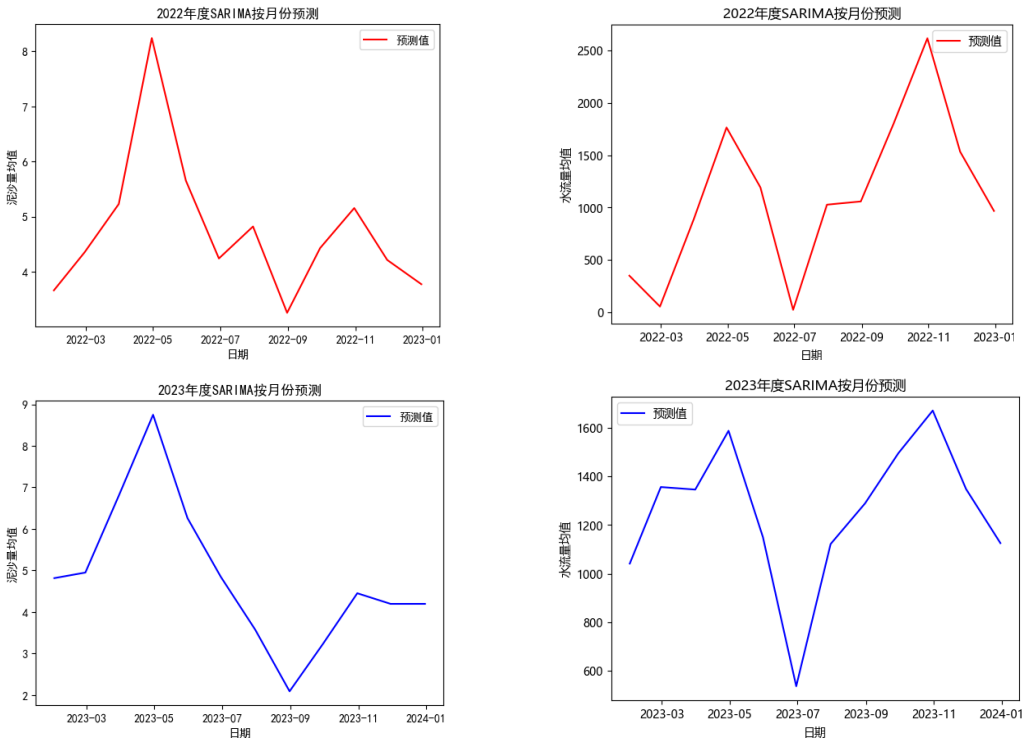

图7 2022-2023未来两年预计和实际的水流量和泥沙量示意图

对于制定最优的采样检测方案，采用增大样本采样的颗粒度，当样本颗粒度增大时，样本的采样频率下降，采样成本下降，采样的连续性下降，预测结果减弱,反之，则增强。流程图如下图8所示：

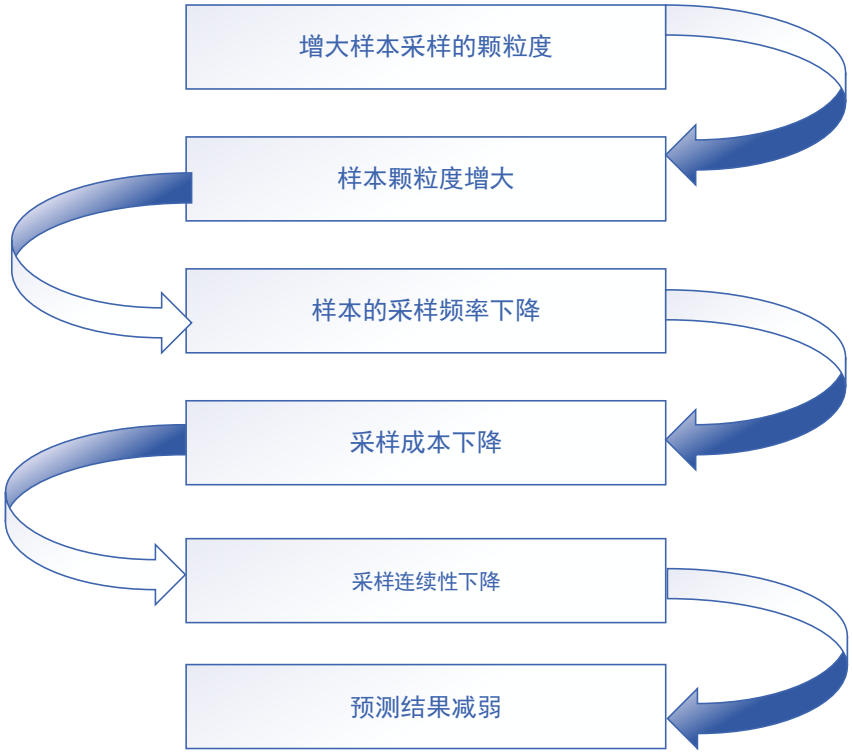

图8 选择最优采样检测方案

对于颗粒度的调节，使用反比例函数进行调节,当自变量x满足函数f(x)：

$$f(x) = \frac{1}{x}, x > 1 \qquad (1-6)$$

取函数f(x)作为颗粒度的调节值，它与采样频率T相关，使得调节后的水沙通量预测采样正确率R和采样频率T以及样本S，以及正确率函数F满足:

$$R > \frac{\sum_{i=1}^{N} F[T * f(x)]}{N} \qquad (1-7)$$

通过计算得到当采样频率T=3h，即8次/天。

问题四：根据水沙通量和河底高程变化，分析每年6-7月“调水调沙”的实际效果。若不“调水调沙”，10年后河底高程如何变。

以2018年调水调沙预计效果和实际效果为例，进行分析。在前三问中已得到水沙通量的变化趋势图，河底高程在年内变化趋势变化较小，在分析不进行调水调沙的预计效果时，采取先将每年6-7月的数据剔除，使用SARIMA模型对6-7月份进行预测分析，得到预测数据后，绘制如下图9所示调水调沙图：

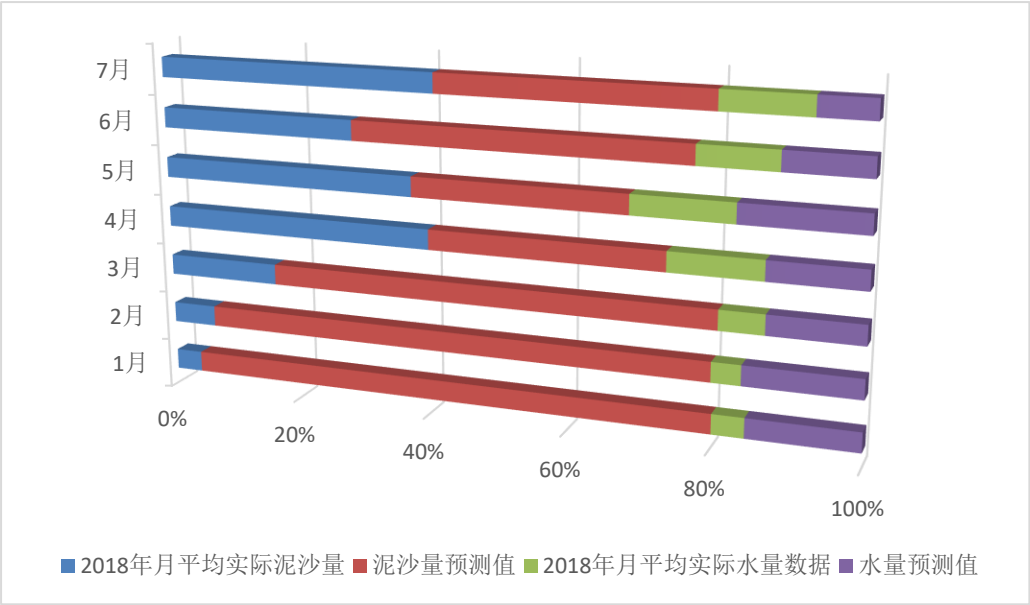

图9　　2018年1-7月份调水调沙图

通过分析数据发现，6-7月调水调沙后的实际水沙通量相对于未调水调沙的水沙通量相对减少，尤其是6月数据中的泥沙量，减少到0.7倍。下图10为非调水调沙年度SARIMA预测的示意图：

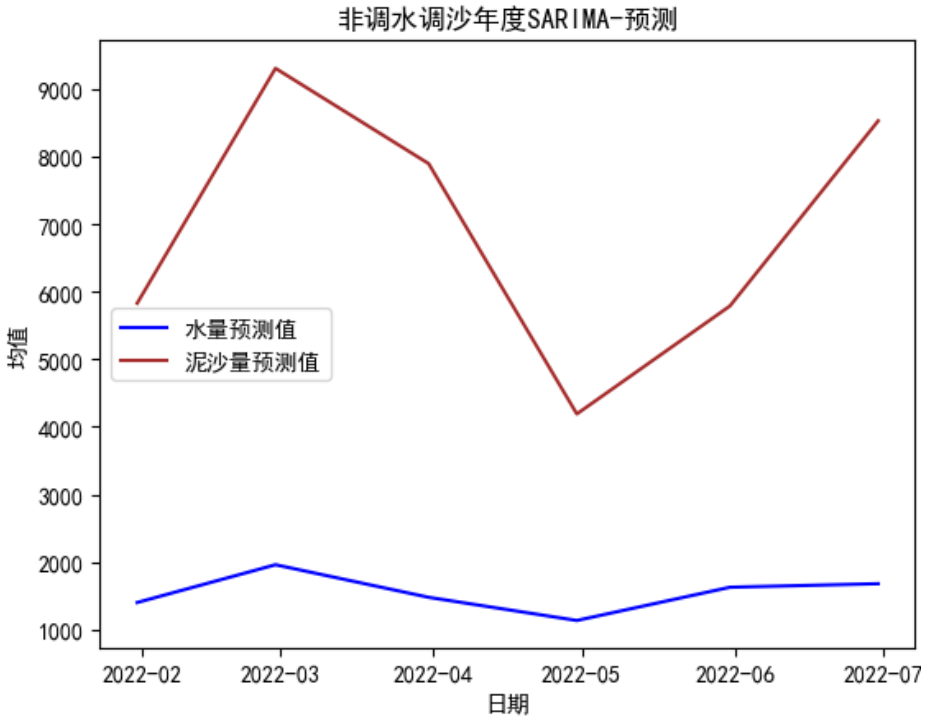

图10　非调水调沙年度SARIMA预测示意图

根据已有数据建立正常年份调水调沙分析图以及不进行调水调沙10年后预测数据值。由于河底高程与河流泥沙淤积量成正相关，经过调水调沙后，河底高程能够基本保持稳定，若不经过调水调沙，河底高程将升高。

五、模型评价

5.1 模型优点

（1）样本使用线性回归进行预测，重采集频率快，得出的数据质量更高

（2）在流量数据弥补缺失值过程中，为防止流量数据重合过高，使用微小变化来使得数据具有一定的波动性，预测得到的数据更合理。

（3）在进行 SARIMA 进行数据预测时，进行参数搜索来预测，并按照季节的滞后性进行季节性划分。

5.2 模型缺点

（1）在进行季节性差分时，得出的 p-value 值在 0.1 左右，对于常规的 0.05 值偏大。

（2）模型对于线性相关性没有进行方差检验。

六、参考文献

[1] 卓金武, MATLAB 数学建模方法与实践, 北京航空航天大学出版社, 2018.

[2] 胡衍坤, 王宁, 刘枢, 姜秋俚, 张楠. 时间序列模型和 LSTM 模型在水质预测中的应用研究[J]. 小型微型计算机系统, 2021,42(08):1569-1573.

[3] 邵鹏郡. 基于 ARIMA 时间序列模型的美国失业率预测研究[J]. 国际公关, 2020(12):395-396.

[4] 杜懿, 麻荣永. 不同改进的 ARIMA 模型在水文时间序列预测中的应用[J]. 水力发电, 2018,44(04):12-14+28.

[5] 葛娜, 孙连英, 赵平, 万莹. 基于 ARIMA 时间序列模型的销售量预测分析[J]. 北京联合大学学报, 2018,32(04):27-33.

附录

1.绘制 ACL 图 python 程序

```
import pandas as pd#导入各种包
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm


# 1. 准备时间序列数据
def getdf():
    datapath = r"C:\Users\Desktop\按小时预测流量泥沙量 2016-2021.xlsx"
    df = pd.read_excel(datapath,usecols=['date','ll'],
parse_dates=True,sheet_name=str(2016),index_col='date')
    return df
def getval():
    df2=pd.DataFrame()
    for year in range(2016,2022):
        df =
pd.read_excel(datapath,usecols=['date','ll'],sheet_name=str(year),index_col='date')
        df2 = pd.concat([df2,df],axis=0)
    return df2
df = getdf()
print(df)
# 2. 绘制 ACF 图
plt.figure(figsize=(12, 6))
sm.graphics.tsa.plot_acf(df['ll'], lags=40, alpha=0.005)
plt.xlabel('Lag')
plt.ylabel('ACF')
plt.title('Autocorrelation Function (ACF)')
plt.show()
```

2.SARIMA 模型算法 python 程序

```
import pmdarima as pm#导入各种包
import matplotlib.pyplot as plt
import pandas as pd
from matplotlib.font_manager import FontProperties
plt.rcParams['font.sans-serif'] = ['Microsoft YaHei']
datapath = r"C:\Users\17192\Desktop\按小时预测流量泥沙量 2016-2021.xlsx"
df = pd.read_excel(datapath,usecols=['date','ll'],
parse_dates=True,sheet_name=str(2018),index_col='date')

df2=pd.DataFrame()
for year in range(2018,2022):
    df =
pd.read_excel(datapath,usecols=['date','ll'],sheet_name=str(year),index_col='date')
    df2 = pd.concat([df2,df],axis=0)
df2 = df2.resample('M').mean()
df2.reset_index()
model = pm.auto_arima(df2['ll'],m=4, stepwise=False)
forecast = model.predict(n_periods=12)
plt.plot(forecast, label='预测值', color='red')
plt.xlabel('日期')
plt.ylabel('水流量均值')
plt.legend()
plt.title('2022 年度 SARIMA 按月份预测')
plt.show()

df2 = pd.concat([df2['ll'],forecast],axis=0)
model = pm.auto_arima(df2,m=4, stepwise=False)
forecast = model.predict(n_periods=12)
```

```
plt.plot(forecast, label='预测值', color='blue')
plt.xlabel('日期')
plt.ylabel('水流量均值')
plt.legend()
plt.title('2023 年度 SARIMA 按月份预测')
plt.show()
```

3.线性回归算法 python 程序

```
import pandas as pd#导入各种包
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
datapath = r"C:\Users\17192\Desktop\附件 1.xlsx";
def gr(res__A):
    mr={}
    for n, s in res__A.groupby(res__A.index.to_period('M')):
        mr[n] = s.resample('H').mean()
    return mr

def getval(i):
    datas__1 = pd.read_excel(datapath,sheet_name=str(i))[['年','月','日','时间','流量(m3/s)','含沙量(kg/m3) ']]
    datas__1[['年','月','日']]=datas__1[['年','月','日']].fillna(method='ffill').astype(int)
    datas__1['date']=pd.to_datetime(datas__1['年'].astype(str) + '-' + datas__1['月'].astype(str) + '-' + datas__1['日'].astype(str)+' ' + datas__1['时间'].astype(str), format='%Y-%m-%d %H:%M')#时间处理
    del datas__1['年']
    del datas__1['时间']
```

```
    datas__1=deletes(datas__1).set_index('date')
    datas__1.index = pd.to_datetime(datas__1.index)
    datas__1=gr(datas__1)
    return datas__1
lis2 = np.array([[]])
df = pd.DataFrame()
excels = pd.ExcelWriter('预测水流量 2016-2021.xlsx',engine='xlsxwriter')
for l in range(2016,2022):
    df = pd.DataFrame()
    d = getval(l)
    for i,j in d.items():
        j=j['流量(m3/s)'].fillna(method="ffill")
        df= pd.concat([df,pd.DataFrame(j.values)],axis=0)
    df.to_excel(excels,sheet_name="{}".format(l),index=False)
excels.save()#保存 excecl
```

4. Mann-kendall 算法 python 程序

```
import pandas as pd
import numpy as np
import pymannkendall as mk_test
import seaborn as sns
import matplotlib.pyplot as plt
import random
datapath = r"C:\Users\17192\Desktop\2016.CSV"
ys = pd.read_csv(datapath,usecols=['date','nsl'])
ys['trend'] = 0  # 创建一个新列用于标记趋势
j=0
for i in range(1, len(ys)):
    # 选择当前日期之前的所有数据点
    subset = ys.iloc[j:i + 1]
    j+=1
```

```
    if j>=10:
        p=random.randint(1,15)
        if p==5:
            j-=1
    if(j in [1000,2000,3000,4000,5000,6000,7000,8000]):
        print(j)

    # 运行 Mann-Kendall 检验
    result = mk_test.original_test(subset['nsl'])

    # 如果检验结果表明存在显著的趋势
    if result[1] <= 0.05:
        ys.at[ys.index[i], 'trend'] = 1  # 将标记设置为 1 表示存在趋势
# 创建 Seaborn 图表
plt.figure(figsize=(12, 6))
sns.lineplot(data=ys, x='date', y='nsl', label='Time Series Data')

# 标记突变值
markers = ys[ys['trend'] == 1]
sns.scatterplot(data=markers, x='date', y='nsl', color='red', marker='o', label='Trend
Change')

plt.xlabel('日期')
plt.ylabel('nsl')
plt.title('Time Series with Trend Changes')
plt.legend()
plt.show()
plt.savefig('pic.png')
```