基于ARIMA模型与 2σ 准则的流量数据异常值检测问题研究

摘 要

如何高效地检测供水系统是否出现漏水是智能水务系统建设待解决的问题，电磁流量计可监测水流量，观测水流量的潜在状态。本文通过建立数学模型，就题目给出的数据集和流量异常检测问题进行分析，给出最优的识别流量异常模型。

针对任务一，对数据模式进行分析，建立识别异常值标准。处理数据集中的重复值和缺失值，调用matplotlib库将清洗后的数据进行可视化，观察到8个地区的序列数据普遍呈现一定趋势性，所以本文决定采用平稳性检验，即ADF检测法，对时间序列进行初步平稳性判断，如果数据平稳则进行下一步的白噪声检验，若非平稳则进行差分将序列变为平稳数据，再进行白噪声检验。对于异常值的标准，根据数据分析结果，本文决定使用ARIMA模型对8个时间序列进行拟合，使用2σ 准则作为异常值判断标准，若残差不在区间内则判定为异常值。

针对任务二，开发一个通用模型。介绍了时间序列分析方法中的ARIMA模型，并探讨了该模型的基本思想和参数选择方法。模型参数p、d、q通过自相关和偏相关函数图像中曲线与置信区间的交点求得，参数选择对于模型的准确性和可靠性非常重要。此外还介绍了2σ准则，该方法可用于检测数据中的异常值。

针对任务三，对模型进行求解、优化和异常值分析。首先对8个区域的数据进行了ADF检验，以确认数据的平稳性。检验得region\_1和region\_7为平稳，其余为非平稳，对非平稳数据进行差分，得到平稳数据。接着，使用Ljung-Box检验对数据进行白噪声检验，确定数据为非白噪声序列。结合ACF图与PACF图，通过比对ADF统计量与临界值的大小，求出参数的依次的值，再对参数进行组合，可以得到ARIMA模型的最佳参数组合为 [(2, 0, 3), (2, 1, 1), (0, 1, 2), (2, 1, 1), (2, 1, 0), (2, 1, 1), (2, 0, 2), (0, 1, 0)]。通过拟合曲线的拟合值和原始值作差得到残差，最后通过2σ准则判断残差是否在区间（μ - 2σ，μ + 2σ），即可以求出异常值。此外，对模型进行优化，采用箱线图与ARIMA模型相结合可以更准确地检测出异常值。最后，分析了异常数据可能产生的原因，为改进检测水流量的采集方法提供了参考。

最后，对模型进行检验和评价。本研究采用残差序列的正态性检验对模型进行检验。通过绘制残差的概率图并进行分析，发现数据点大致分布在直线附近，说明残差近似服从正态分布，这表明该模型对数据的拟合效果良好。因此，本研究得出的模型预测结果较为可靠，可以较好地反映实际情况。

关键词：ARIMA模型 ADF检测法 2σ准则 箱线图

目 录

[一、问题重述 3](#_Toc100469970)

[1.1问题背景 3](#_Toc100469971)

[1.2问题的数据条件 3](#_Toc100469972)

[1.3问题的提出 3](#_Toc100469973)

[二、问题分析 3](#_Toc100469974)

[2.1问题总分析 3](#_Toc100469975)

[2.2具体问题分析 3](#_Toc100469976)

[2.2.1任务一的分析 3](#_Toc100469977)

[2.2.2任务二的分析 3](#_Toc100469978)

[2.2.3任务三的分析 3](#_Toc100469979)

[三、模型假设 4](#_Toc100469981)

[四、符号说明 4](#_Toc100469982)

[五、模型建立与求解 4](#_Toc100469983)

**[5.1](#_Toc100469984)**[任务一求解 4](#_Toc100469984)

[5.1.1数据模式分析 4](#_Toc100469985)

[5.1.2检测异常标准建立 5](#_Toc100469986)

**[5.2](#_Toc100469988)**[任务二的模型 5](#_Toc100469988)

[5.2.1 ARIMA模型 5](#_Toc100469989)

[5.2.2模型参数确定 5](#_Toc100469990)

[5.2.3](#_Toc100469991) *[2σ](#_Toc100469991)*[准则 6](#_Toc100469991)

**[5.3](#_Toc100469992)**[任务三求解 7](#_Toc100469992)

[5.3.1模型求解 7](#_Toc100469993)

[5.3.2模型优化 12](#_Toc100469994)

[5.3.3异常值分析 12](#_Toc100469995)

[六、模型检验 13](#_Toc100469997)

[七、模型评价 14](#_Toc100470000)

[八、结语 14](#_Toc100470001)

[九、参考文献 14](#_Toc100470001)

十、附录.........................................................................................................................................................15

1. **问题重述**

## 1.1问题背景

随着智能水务系统的建设，如何高效地检测供水系统是否出现漏水，成为了一个亟待解决的问题。如何及时发现漏水情况，更好地保障供水质量和供水安全，一直是水务领域研究的重要方向。而电磁流量计作为一种新兴的水流量检测技术，利用电磁感应原理，根据导电流体通过外加磁场时感生的电动势来测量导电流体流量，因此能够对水流量进行监测，并观测出水流量的状态与潜在泄漏，但是如何有效地检测水流量异常仍然是一个挑战。本研究的目的是提出一个数据分析模型，以解决漏水检测问题，为智能水务系统的建设提供技术支持。

## 1.2问题的数据条件

数据文件包含8个不同虚拟区域中输入水流量和输出水流量之差的数据。

## 1.3问题的提出

1. 根据给出题目给出的数据文件，分析数据模式，综合考虑多种异常因素，建立从给定数据中识别流量异常的标准。
2. 根据任务一的数据分析结果，开发一个通用模型对8个虚拟区域进行异常值检测。
3. 在给定数据集上做模型测试，并对建模和异常值检测的结果给出合理解释。

# 二、问题分析

## 2.1问题总分析

首先，对题目给出的数据集进行数据清洗，其次，对时间序列数据进行平稳性和白噪声检验，再次，如果数据为非平稳性数据，则通过差分的方法进行处理，直到该数据为平稳性数据，最后，确定ARIMA模型最佳参数，建立模型，并利用模型残差分布方法检验模型的合理性。

## 2.2具体问题分析

### 2.2.1任务一的分析

任务一要求分析数据模式，并建立识别水流量异常标准。在分析数据模式时，考虑到题目给出了8个虚拟区域输入水流量与输出水流量之差的数据，本文对数据进行可视化处理，研究数据平稳性，对数据预处理后建立了识别异常值标准。

### 2.2.2任务二的分析

任务一已对数据进行预处理，任务二要求建立异常值识别模型。对于模型建立过程中的未知参数，通过数据预处理求得ACF图像与PACF图像，并求解得到最优参数应用于后续解题过程 。

### 2.2.3任务三的分析

在任务一与任务二的基础上，在题目给出的数据集上做模型测试，对模型进行分析，对异常值检测结果做出合理解释。

# 

# 三、模型假设

1. 假设题目给出的数据是可靠的
2. 假设流量数据在短时间内是稳定的，即短时间内不会发生大幅度的变化，可以使用统计方法对数据进行分析
3. 假设流体的电导率是稳定的
4. 假设流量计准确度已经达到一定精度，因此不考虑流量计本身的误差对数据分析的影响
5. 假设在模型开发过程中仅考虑题目中已涉及的主要条件，即异常值的出现是由于系统故障引起的漏水，不考虑水管材质，水压变化等因素
6. 假设所建立的模型可以泛化到其他水管网系统中，而不仅仅局限于本题提供的水管网系统

# 符号说明

|  |  |  |
| --- | --- | --- |
| **符号** | **说明** | **单位** |
| *p* | 预测模型中采用的时间序列数据本身的滞后数 | 无 |
| *q* | 预测模型中采用的预测误差的滞后数 | 无 |
| *Фi* | 自回归参数 | 无 |
| *θi* | 移动平均的参数 | 无 |
| *εt* | 误差项，通常是独立的 | 无 |
| *Xt* | 时间序列数据 | 无 |
| *d* | 时间序列数据需要进行差分的次数，*d*∈*Z*，*d >* 0 | 无 |
| *L* | 滞后算子 | 无 |

注：表中未提及的符号在文中出现时进行说明。

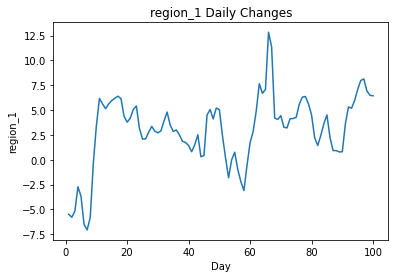
**五、模型的建立与求解**

## 5.1任务一求解

### 5.1.1数据模式分析

（1）清洗数据中重复出现的值，直接将重复值删去；

（2）处理数据的缺失值，使用该日前后2d 的平均值代替；

（3）调用matplotlib库将清洗后的数据进行可视化：

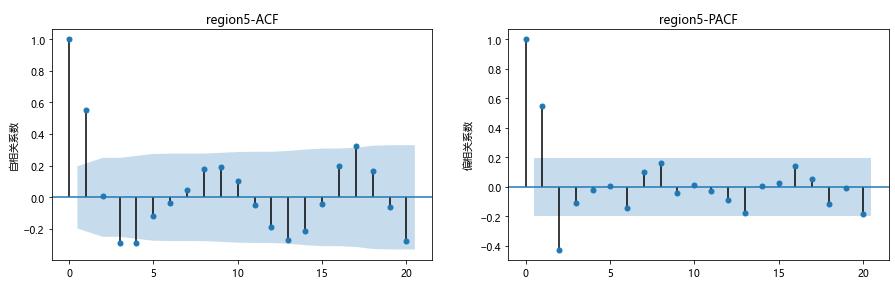
以分析region\_1为例，从图中可以看出，region\_1时间序列的数值观测值呈现出一定的趋势性，即整体呈现出逐渐上升的趋势，但波动幅度较大。此外，还可以看出在时间序列的前期部分，波动幅度较大，而后期部分则呈现出相对平稳的状态。

针对这样的时间序列，本文采用经典的时间序列分析方法，即平稳性检验。

（4）采用ADF检测法判断，对时间序列数据进行初步平稳性判断，如果数据为平稳性数据，则进行下一步，即白噪声检验，如果序列为非平稳性数据，则通过差分的方法进行处理，直到该序列为平稳性数据，再进行白噪声检验。

下面是region\_5的ACF图和PACF图：以region\_5为例，通过观察ACF和PACF图像，可以得到结论：ACF的第一个峰（滞后1）明显高于其他滞后，这表明序列存在较强的自相关性；PACF的第一个峰（滞后1）也很高，这也表明序列存在较强的自相关性。

通过ACF和PACF的初步分析，可以得出这个序列不是平稳的。因为如果数据是平稳的，就不会存在较强的自相关性，并且ACF和PACF的峰值都不应该超过置信区间。



### 5.1.2检测异常标准建立

基于数据模式分析结果，本文采用ARIMA模型，拟合出8个虚拟区的时间序列数据趋势图，使用 2σ 准则判断原始数据与拟合数据残差，如果残差在区间（μ - 2σ，μ + 2σ）内，则判定为正常值，否则为异常值，并结合箱线图识别出其他异常值。

## 5.2任务二的模型

5.2.1 ARIMA模型

ARIMA（差分自回归移动平均）模型是一种经典的时间序列分析方法，广泛应用于预测和建模。ARIMA 模型中，AR 是自回归，MA 为滑动平均。模型的基本思想是用现有的平稳时间序列数据去预测将来的值，即通过现有的稳定时间序列数据求出未来的数据。ARIMA 模型表示为：

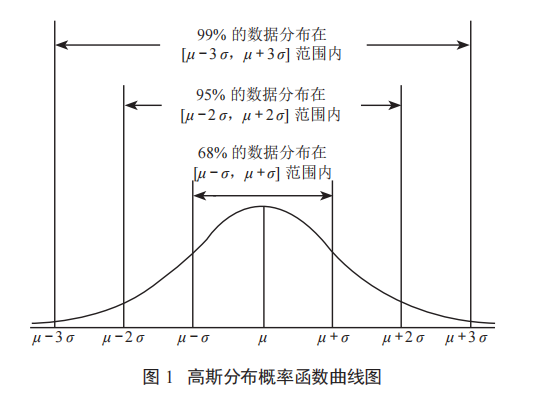
5.2.2 模型参数确定

ARIMA模型的参数通常由三个部分组成：p，d，q。对第一部分数据进行平稳性处理时记录的差分次数即模型参数 d的值，通过自相关和偏相关函数图像中曲线的截尾和拖尾性求出参数 p 和 q 的值。ARIMA模型的参数选择对于模型的准确性和可靠性非常重要

5.2.3 2σ准则

若随机变量 X 服从 1 个位置参数为 μ（均值）、尺度参数为 σ（标准偏差）的概率分布，且其概率密度函数为：

则这个随机变量为正态随机变量，正态随机变量服从正态分布，即高斯分布，记作 X～N（μ，σ2 ）（N 为样本数），其概率密度函数图像如下图 1 所示：



*2σ* 准则是先假设 1 组检测数据只含有随机误差，对其进行计算处理得到 *σ* 和 *μ*，确定 1 个区间 （*μ* - 2*σ*，*μ* + 2*σ*），该组数据在（*μ* - 2*σ*，*μ* + 2*σ*）区间的概率为 95%，即可将超过这个区间的值判定为异常值。

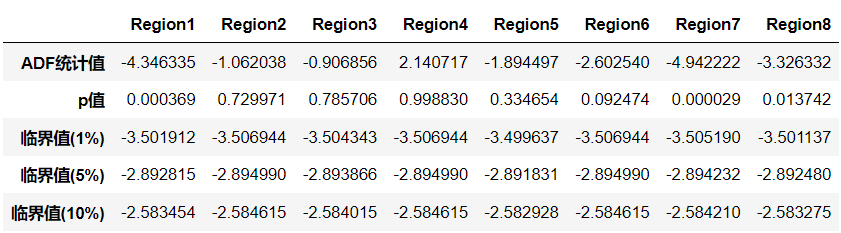
## 5.3任务三求解

5.3.1模型求解

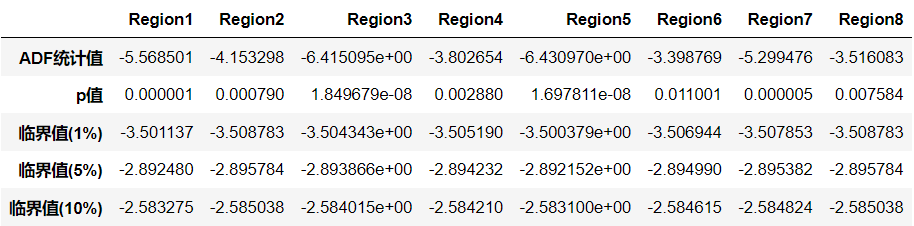
1．ADF检验

对 8个区域数据进行平稳性检测，即 ADF 检测（单位根检验），变量为数据中自带的标签

原始数据的ADF检验结果：



一阶差分后的ADF检验结果：



注：*ADF统计量对应*统计量

*P*为统计量对应的概率值

*临界值* 分别对应 3个置信度（1%，5%，10%）的临界统计值

2. 白噪声检验

ARIMA模型建模的基本前提是数据是平稳的且为非白噪声序列。

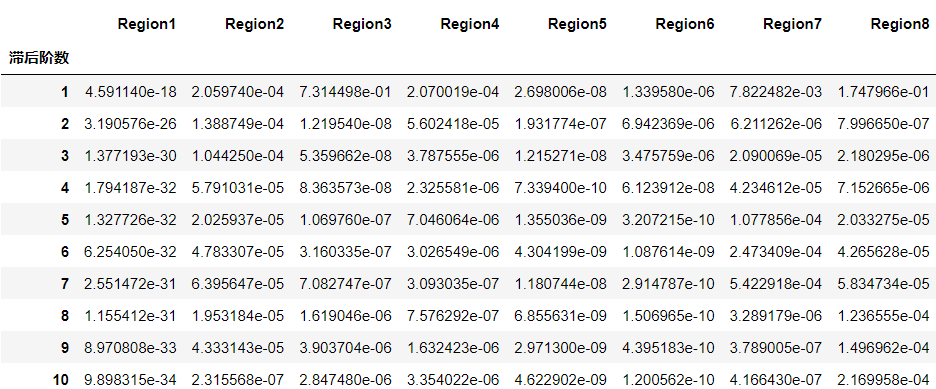
白噪声是一种时间序列，其包含了无规律的波动，这些波动是由随机变量所导致的，并且彼此之间没有相关性。

白噪声检验可以用于检测原始数据是否存在随机性和无规律性。常用的白噪声检验方法有Q统计量检验和Ljung-Box检验。本文采用Ljung-Box检验。

Ljung-Box检验是一种基于ACF的检验方法。该检验统计自相关函数的前几个值之和的平方和，得到Ljung-Box统计量，并进行显著性检验。如果Ljung-Box统计量大于其对应显著性水平下的临界值，则拒绝原假设，即序列不是白噪声序列。

我们对每个地区都进行了Ljung-Box检验，每行结果表示在此滞后期下Ljung-Box检验的p值。

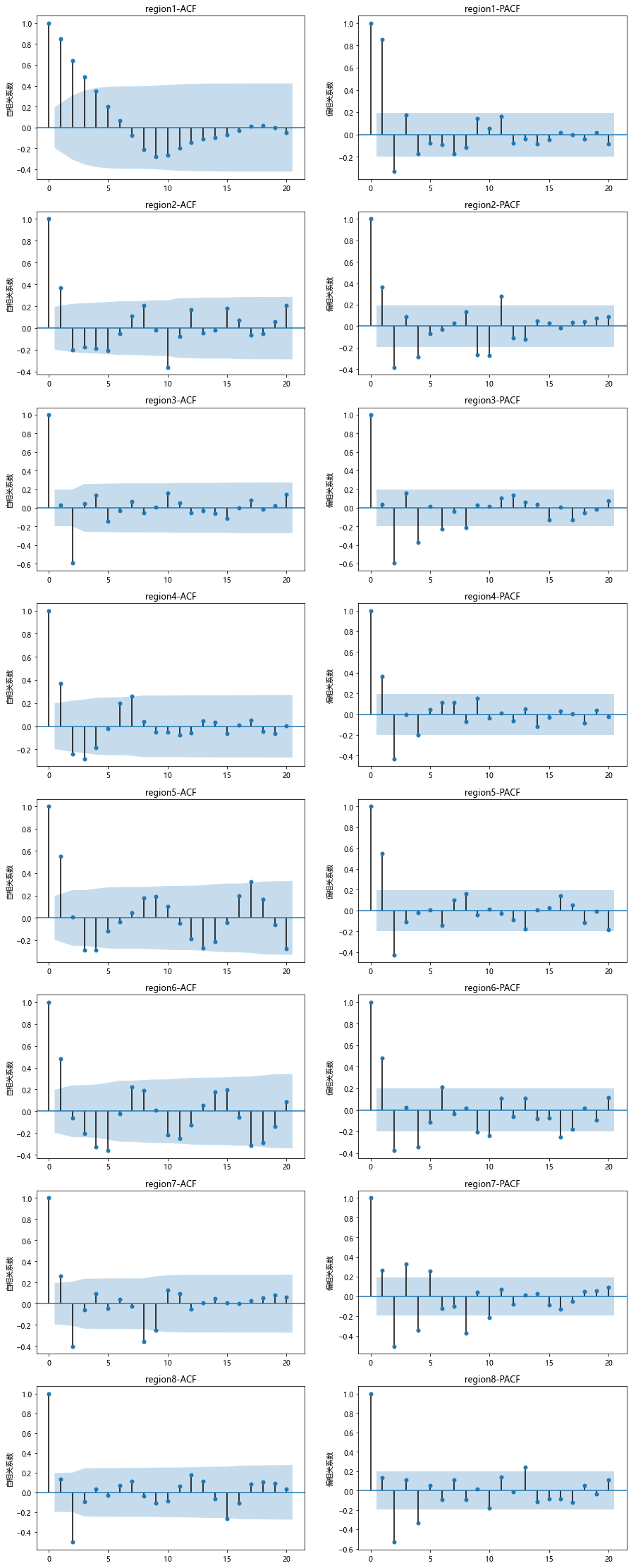
以下为对8个地区的时间序列白噪声检验结果：



根据Ljung-Box检验的原理，如果数据为白噪声序列，那么p值应该大于显著性水平（通常为0.05）。对于每个地区的检验结果， Ljung-Box的p值都很小，这说明8个地区的时间序列都不是白噪声序列，数据之间存在一定的相关性。

3. 模型参数求解

分析数据模式时得到的自相关和偏相关函数图像分别如下图所示（Autocorrelation为自相关函数图像，Partial Autocorrelation为偏相关函数图像，其中，蓝色区域划分是为了在特定的函数相关性取值时取相关系数的值）（以下仅列出前三个地区）：



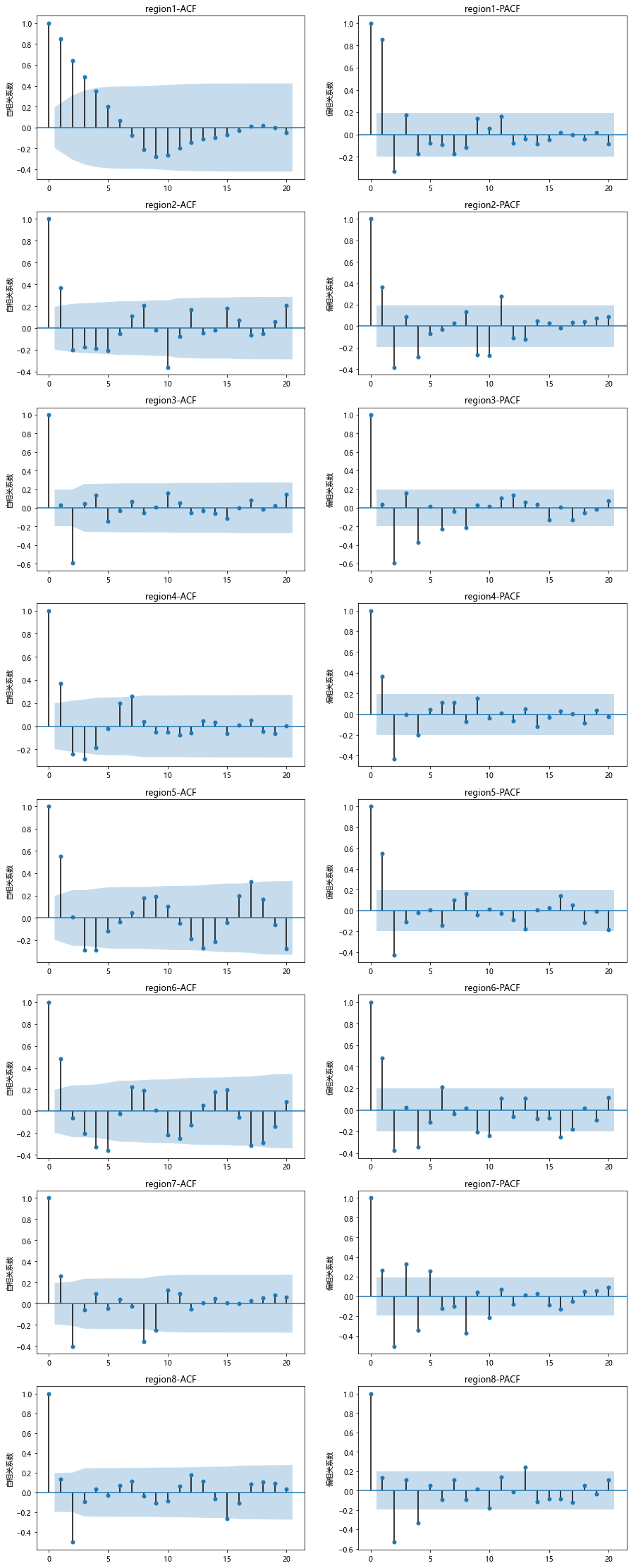
图像截尾和拖尾含义：

截尾：在大于某个常数k后快速趋于0为k阶截尾

拖尾：始终有非零取值，不会在k大于某个常数后就恒等于零(或在0附近随机波动)



以region\_7为例，由图中可以确定两个图都在2阶后截尾，因此可以选定p, q值

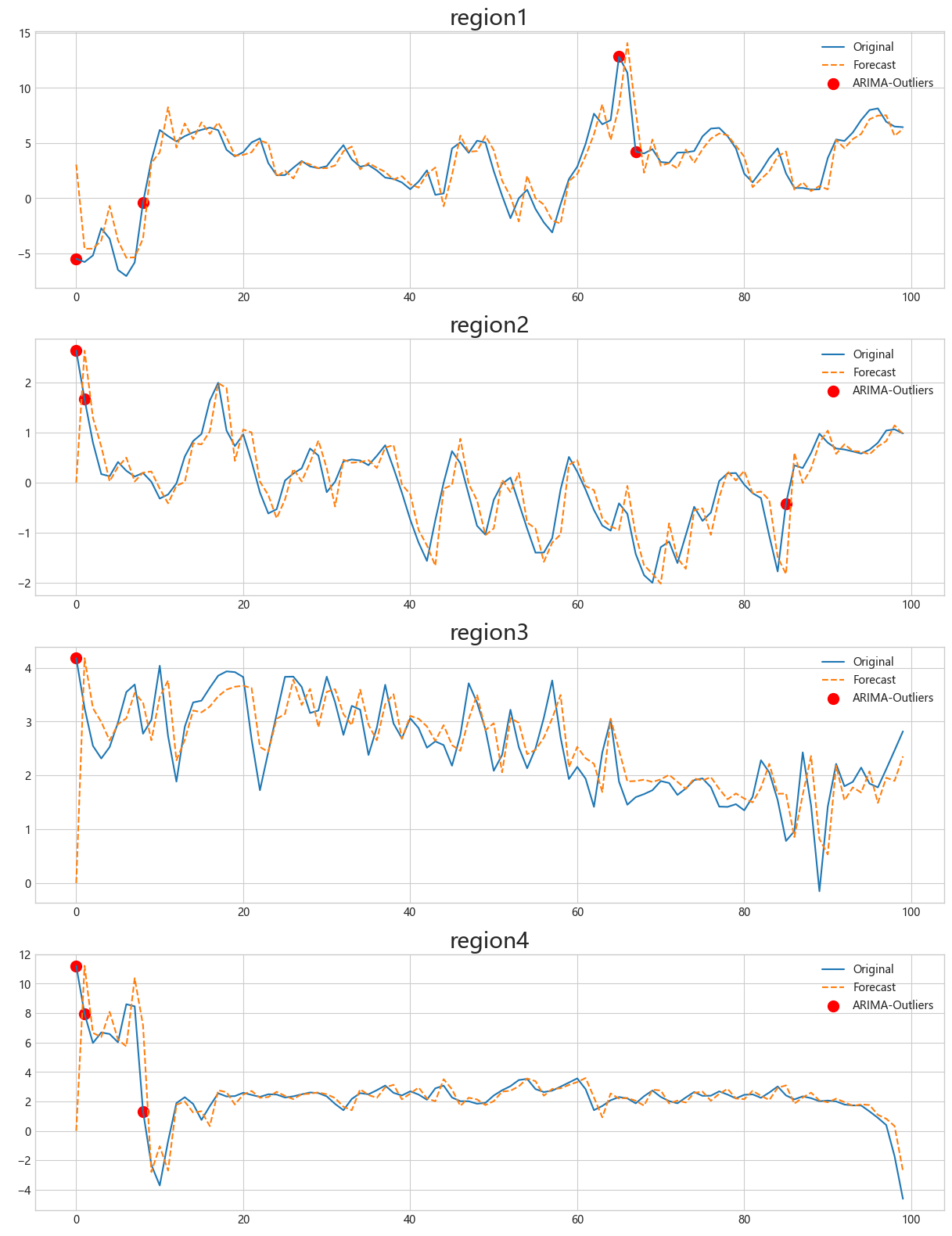


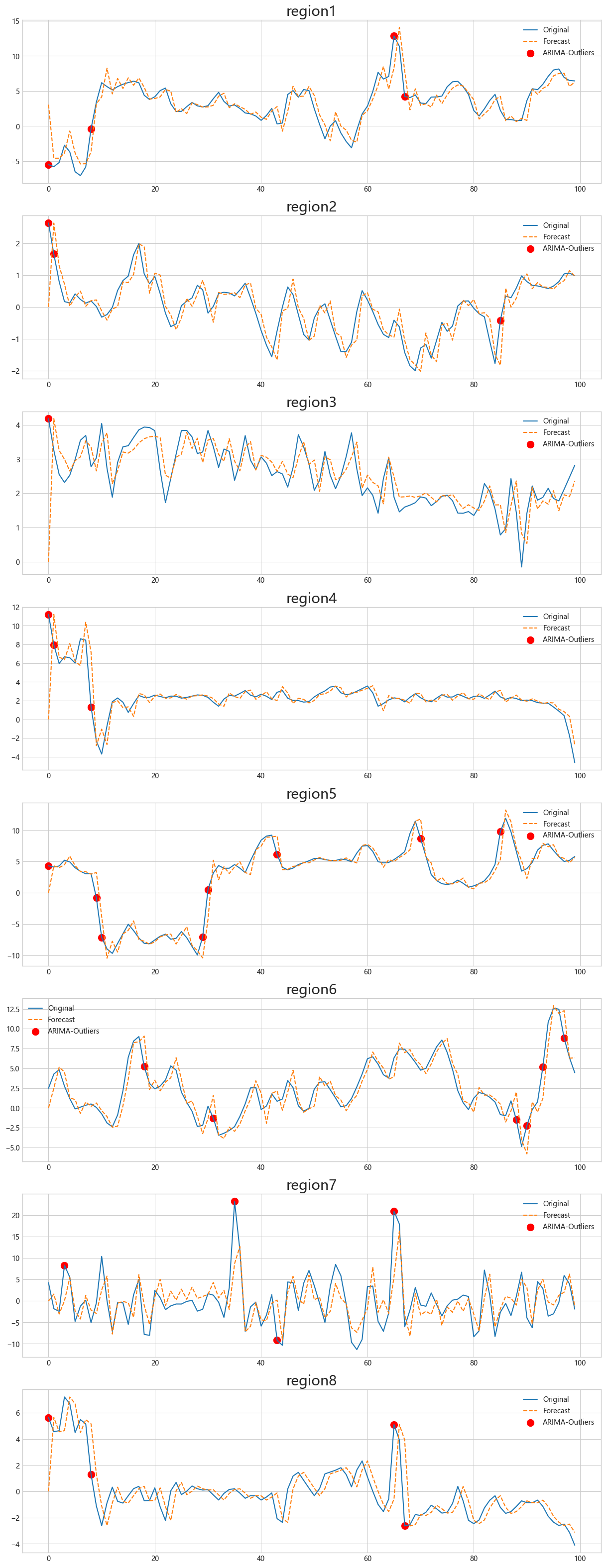
结合ACF图与PACF图，通过比对ADF统计量与临界值的大小，可知各个区域的对应的差分次数依次为0、1、1、1、1、1、0、1。

通过自相关和偏相关函数图像可以求出每个地区对应的AR模型的 p 值依次为2、2、0、2、2、2、2、0，MA模型的 q 的值依次为3、1、2、1、0、1、2、0。

4. 通过对参数 *p*，*d*，*q* 的组合，可以得到每个区域的ARIMA模型的最佳参数组合依次为 (2, 0, 3), (2, 1, 1), (0, 1, 2), (2, 1, 1), (2, 1, 0), (2, 1, 1), (2, 0, 2), (0, 1, 0)

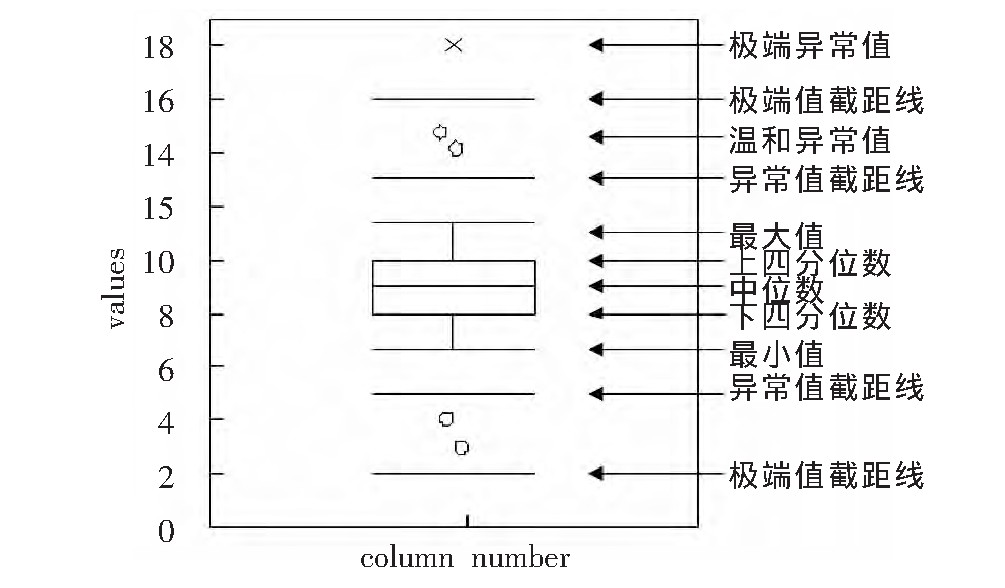
5. 通过 ARIMA 模型中的拟合函数求出该时间序列的拟合趋势，得到的时间序列数据拟合曲线下图所示。从下列各图可以看出预测值和真实值能够较好地拟合在一起，说明 ARIMA 模型效果较好。然后通过拟合曲线中的拟合值和原始值作差得到残差，最后通过 2σ 准则判断残差是否在区间（μ - 2σ，μ + 2σ），即可以求出异常值(红点为异常值)：



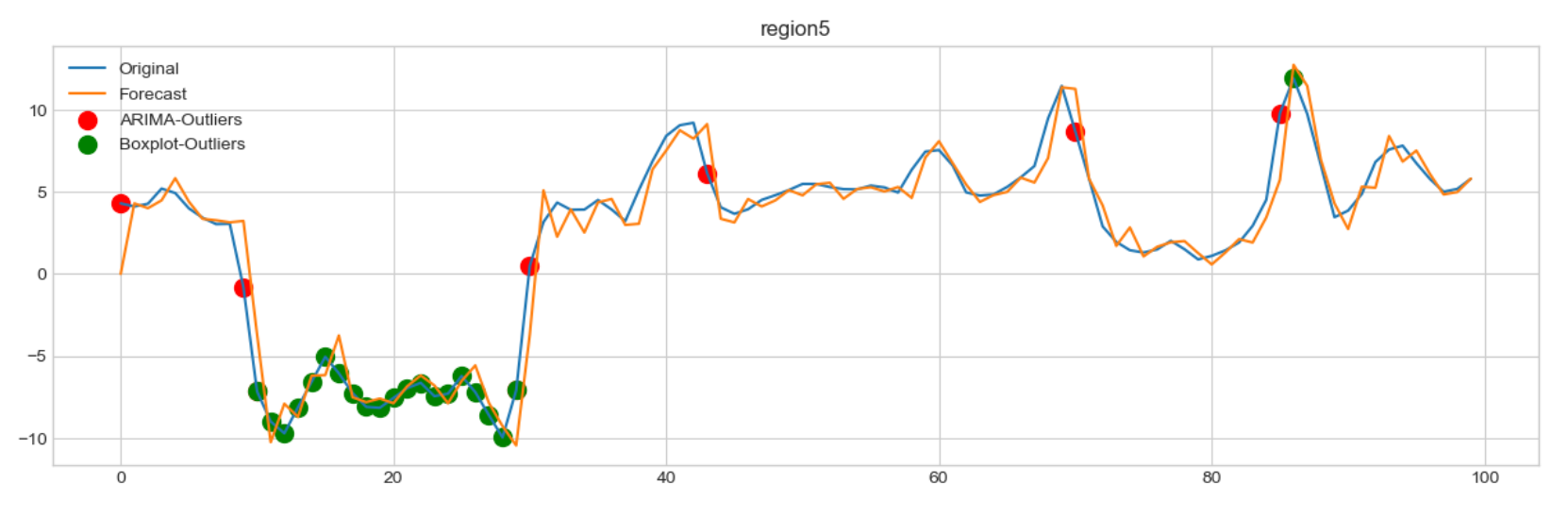


5.3.2模型优化

1. ARIMA模型是基于时间序列的相关性来建模的，因此，它无法检测到时间序列中的局部异常值，因为这些异常值会破坏时间序列的自相关性。为更加准确地检测出异常值，本文结合箱线图找出异常点

2. 原理：箱形图中心位置为中位数,箱子的长度表示四分位数的间距(IQR),两端分别是上四分位数(Q3)和下四分位数(Q1),箱两端的须为最大值和最小值。箱形图法定义的异常值是指样本数据中大于Q3+1.5IQR和小于Q1-1.5×IQR。其中Q3+1.5×IQR和Q1-1.5IQR为异常值截距线,Q3+3IQR和Q1-3IQR为极端值截距线,介于异常值截距线与极端值截距线之间的异常值叫温和异常值,位于极端值截距线以外的异常值叫极端异常值

3.接着用箱线图呈现数据的分布情况和异常值，对比上面的异常值检测结果，发现region\_5有较多局部异常值未被检测出



4.因此，模型检测得到的异常点即为红色点与绿色点之和

5.3.3异常值原因分析

日流量数据的异常值可能导致流量数据的分析出现错误，影响流量检测点判断异常决策，以及水流量数据采集整体和数据等质量评价工作的效率。因此，对于异常值的判断可以很大程度地提高水的使用效率，通过判断流量异常的原因，可改进检测水流量的采集方法。

异常数据可能产生的原因很多，经过分析可能的原因如下：

1）检测水流量设备的传感器出现故障；

2）流量计设备出现老化，导致取值结果出现较大的误差；

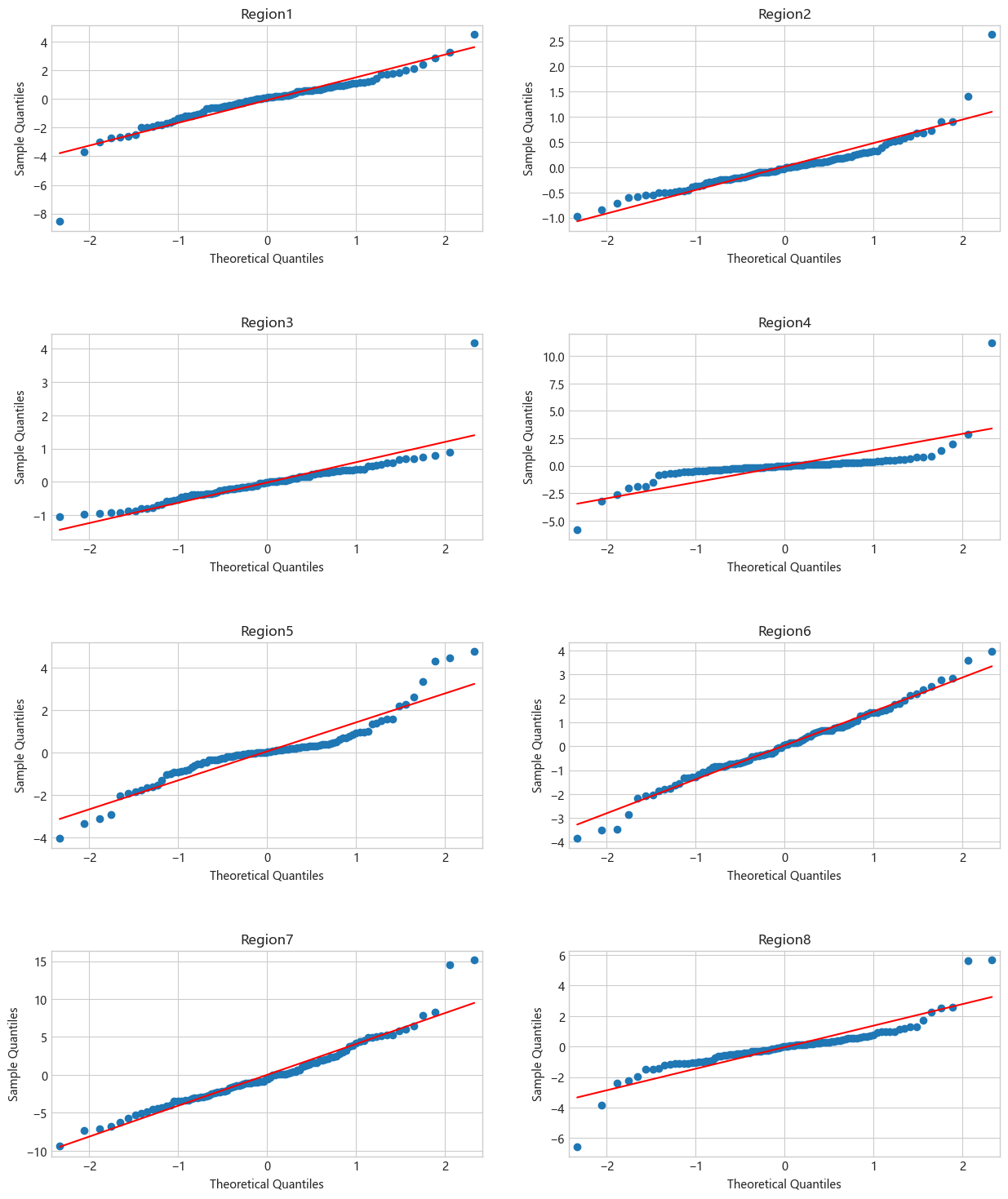
3）检测水流量的操作员出现操作失误；

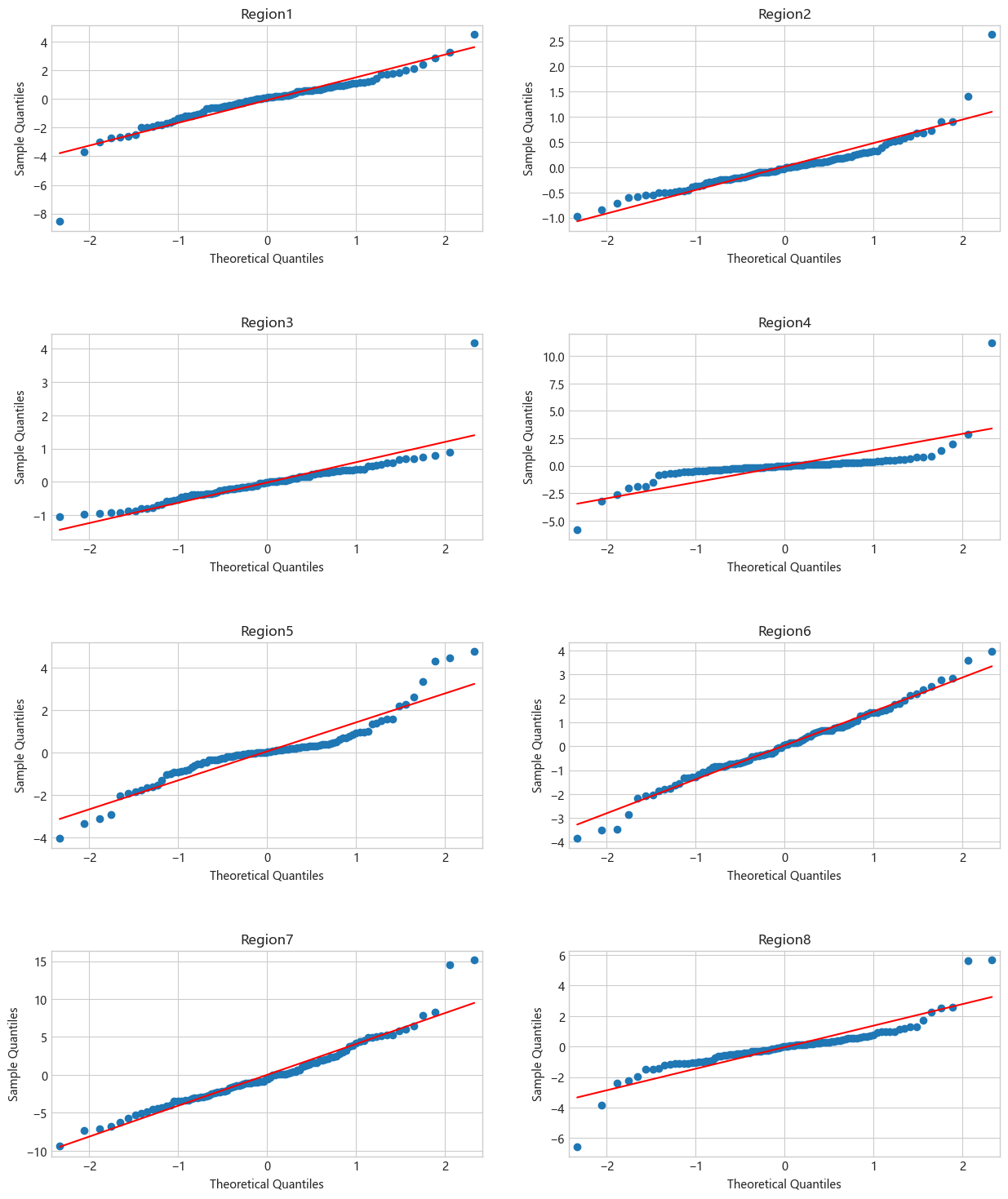
4）出现自然灾害等非人为因素，导致检测水流量设备严重不符合流量的日常值。

通过对这些原因的分析，每个流量检测点可以针对产生异常值的原因进行改进，从而提高水流量监测数据的质量。

# **六、模型的检验**

本研究采用残差序列的正态性检验对模型进行检验，首先绘制残差的概率图：





对概率图进行分析，发现数据点大致分布在直线附近，残差近似服从正态分布，说明该模型对数据的拟合效果良好，并且模型的预测结果能够较为可靠地反映实际情况。

因为正态分布是一种常见的概率分布，其具有较好的数学性质和统计性质，因此，如果残差近似服从正态分布，就意味着残差的变化是随机的、无规律的，不会对模型的预测结果产生显著的影响。

**七、模型评价**

这个数学模型的优势之一就是能够更有效地检测水流的异常情况。当电磁流量计检测到任何水流异常，便可以立即将信号传送到供水系统，提醒及警告人们尽快修理水管或供水系统，从而可以防止漏水问题恶化。

可是，我们的数学模型也有一个局限。由于每个区域的人口众多，在较小的水管分支里，水输入和输出量差异不大，因此我们无法检测到当中异常的水流，也就忽略了较小的水管分支中的漏水问题。

**八、结语**

总括而言，我们建立了一个能够同时了解多个流量计的数学模型及有效地检测流量异常的情况，并 (1)分析了数据模式并找出识别数据异常的标准，(2)建立了通用的数学模型，根据数据分析的结果对8个虚拟区域进行了异常水流量的检测，以及 (3)根据数据测试我们建设的数学模型，然后就模型和异常检测的结果做出解释。我们希望通过提醒供水机构和大众关于漏水的情况，防止潜在故障恶化，避免公众面临的严重漏水问题。

**九、参考文献**

[1]李玉梅.数据的正态性检验方法[J].怀化学院学报,2015,34(11):81-82.

[2]莫达隆.利用ADF检验对时间序列进行建模[J].时代金融,2010,(04):46-48.

[3]张顺先,邱磊,张少言,李翠,胡骏,田黎明,鹿振辉.ARIMA模型的建立及对中国肺结 核月报告例数的预测效果研究[J].中国防痨杂志,2020,42(06):614-620.

[4]赵和松,王圆圆,孙爱民.一种基于ARIMA模型与3σ准则的取水异常检测方法[J].水 利信息化,2022,(01):35

[5]于宁莉,易东云,涂先勤.时间序列中自相关与偏相关函数分析[J].数学理论与应 用,2007,(01):54-57.

[6]李梦辉,桑小田,田振川,申献夫.箱形图在长江流域棉区棉花含杂率异常值检验中的 应用[J].湖北农业科学,2016,55(11):2895-2898+2954.

# 十、附录

|  |
| --- |
| 附录1 |
| 介绍：支撑材料的文件列表 |
| 数据文件.xls  data\_cleaning.py  ACF和PACF图像.py  ADF检验.py  白噪声检验.py  fit\_ARIMA.py  概率图.py  异常点检测.py  模型效果图.jpg  ACF和PACF图.jpg |

附录2主要代码

任务一：数据清洗 （文件名：data\_cleaning.py）

# 清洗数据

import pandas as pd  
def clean(df):  
 df.drop\_duplicates(inplace=True) # 去除重复行  
 # 使用相邻两个数的平均值填充缺失值  
 for col in df.columns:  
 if df[col].isnull().sum() > 0: # 如果列中有缺失值  
 values = df[col].values # 获取列数据  
 for i in range(len(values)):  
 if pd.isna(values[i]): # 找到缺失值  
 if i == 0: # 缺失值在列表开头，用后一个数据填充  
 avg = (values[i+1] + values[i+2]) / 2  
 elif i == len(values)-1: # 如果缺失值在列表结尾，用前一个数据填充  
 avg = (values[i-2] + values[i-1]) / 2  
 else: # 缺失值在列表中间，用前后两个数据的平均值填充  
 avg = (values[i-1] + values[i+1]) / 2  
 if pd.isna(avg): # 如果前后两个数据都是缺失值，则用列的均值填充  
 values[i] = df[col].mean()  
 else:  
 values[i] = avg  
 df[col] = values # 将填充好的数据放回数据集  
 return df

任务三: (1)模型求解 (文件名：ADF检验.py)

import pandas as pd

from data\_cleaning import clean

from statsmodels.tsa.stattools import adfuller

data = pd.read\_excel('数据文件.xls')

data = clean(data)

index = ['ADF统计值', 'p值', '临界值(1%)', '临界值(5%)', '临界值(10%)']

df\_0 = pd.DataFrame(index=index, columns=['Region%d' % i for i in range(1, 9)])

df\_1 = pd.DataFrame(index=index, columns=['Region%d' % i for i in range(1, 9)])

data = pd.read\_excel('数据文件.xls')  
data = clean(data)

for i in range(1, 9):

ts = data.iloc[:, i]

# ADF检验数据

# 没有差分的DataFrame

result = adfuller(ts)

df\_0['Region%d' % i] = [result[0], result[1], result[4]['1%'], result[4]['5%'], result[4]['10%']]

# 进行差分的DataFrame

ts = ts.diff().dropna()

result = adfuller(ts)

df\_1['Region%d' % i] = [result[0], result[1], result[4]['1%'], result[4]['5%'], result[4]['10%']]

print('原始数据的ADF检验结果：')

print(df\_0)

print('一阶差分后的ADF检验结果：')

print(df\_1)

任务三: (1)模型求解 (文件名：白噪声检验.py)

from statsmodels.stats.diagnostic import acorr\_ljungbox  
import pandas as pd  
from data\_cleaning import clean  
  
data = pd.read\_excel('数据文件.xls')  
data = clean(data)  
  
# diff\_ 标记为1的数据需要进行一阶差分  
diff\_ = [0, 1, 1, 1, 1, 1, 0, 1]  
df = pd.DataFrame(index=[i for i in range(1, 11)], columns=['Region%d' % i for i in range(1, 9)])  
for i in range(1, 9):  
 ts = data.iloc[:, i]  
 if diff\_[i - 1] == 1:  
 ts = ts.diff().dropna()  
 # 进行Ljung-Box检验  
 result = acorr\_ljungbox(ts, lags=10)  
 df['Region%d' % i] = result['lb\_pvalue']  
df.index.name = '滞后阶数'  
print(df)

任务三: (1)模型求解 (文件名：ACF和PACF图像.py)

from data\_cleaning import clean  
import matplotlib.pyplot as plt  
from statsmodels.graphics.tsaplots import plot\_acf, plot\_pacf  
plt.rcParams['font.sans-serif'] = ['Microsoft YaHei']  
data = pd.read\_excel('数据文件.xls')

data = clean(data)

fig, ax = plt.subplots(8, 2, figsize=(15, 40))  
  
# diff\_ 标记为1的数据需要进行一阶差分  
diff\_ = [0, 1, 1, 1, 1, 1, 0, 1]  
  
for i in range(1, 9):  
  
 # 绘制时间序列的ACF和PACF图  
 # 图表每一行代表一个地区  
 ts = data.iloc[:, i]  
 if diff\_[i - 1] == 1:  
 ts = ts.diff().dropna()  
  
 plot\_acf(ts, ax=ax[i-1, 0])  
 plot\_pacf(ts, ax=ax[i-1, 1], method='ywm')  
 ax[i-1, 0].set\_title('region%d-ACF' % i)  
 ax[i-1, 1].set\_title('region%d-PACF' % i)  
 ax[i-1, 0].set\_ylabel('自相关系数')  
 ax[i-1, 1].set\_ylabel('偏自相关系数')  
  
plt.show()

任务三: (1)模型求解 (文件名：fit\_ARIMA.py)

# 拟合ARIMA模型  
from statsmodels.tsa.arima.model import ARIMA  
import warnings  
warnings.filterwarnings('ignore', message="Non-invertible starting MA parameters found")  
  
def main():  
 # 模型的参数(p, d, q), 依次对应一个地区的模型参数

param = [(2, 0, 3), (2, 1, 1), (0, 1, 2), (2, 1, 1), (2, 1, 1), (2, 1, 1), (2, 0, 2), (0, 1, 0)]  
 resid, forecast = [], []  
 for i in range(1, 9):  
 ts = data.iloc[:, i]  
 # 训练 ARIMA 模型  
 model = ARIMA(ts, order=param[i - 1])  
 result = model.fit()  
 # 残差  
 resid.append(result.resid)  
 # 预测  
 forecast.append(result.predict(start=0, end=len(ts)-1, typ='levels'))  
  
 return forecast, resid

任务三: (2)模型优化 (文件名：异常点检测.py)

from data\_cleaning import clean  
import fit\_ARIMA  
import numpy as np

data = pd.read\_excel('数据文件.xls')

data = clean(data)  
forecast, resid = fit\_ARIMA.main()  
fig, ax = plt.subplots(8, 1, figsize=(15, 40))  
for i in range(1, 9):  
  
 ts = data.iloc[:, i]  
  
 ax[i-1].plot(ts, label='Original')  
 ax[i-1].plot(forecast[i-1], linestyle='--', label='Forecast')  
  
 # 因残差近似服从正态分布，所以基于2σ准则判断异常点  
 resid\_std = resid[i-1].std()  
 resid\_mean = resid[i-1].mean()  
  
 # 异常值的布尔索引  
 outliers = (resid\_mean + 2 \* resid\_std < resid[i-1]) | (resid[i-1] < resid\_mean - 2 \* resid\_std)  
  
 # ARIMA模型标注的异常点为红色  
 # 箱形图标注的异常点为绿色，有可能把同一个位置的红色点覆盖掉  
 ax[i-1].scatter(np.arange(100)[outliers], ts[outliers], c='red', s=100, label='ARIMA-Outliers')  
  
 # 箱形图判读异常值  
 q1, q3 = ts.quantile(q=[0.25, 0.75])  
 iqr = q3 - q1  
 outliers\_index = (ts < (q1 - 1.5 \* iqr)) | (ts > (q3 + 1.5 \* iqr))  
 ax[i-1].scatter(np.arange(100)[outliers\_index], ts[outliers\_index], c='green', s=100, label='Boxplot-Outliers')  
 ax[i-1].set\_title('region%d' % i, fontsize=20)  
 ax[i-1].legend()  
  
plt.show()