



河海大学

# 《时间序列分析》

## 课程作业报告

学号姓名\_\_\_\_\_李宗霖\_\_\_\_\_

专业年级\_\_\_\_\_20 级大数据管理与应用\_\_\_\_\_

指导老师\_\_\_\_\_韦庆明\_\_\_\_\_

报告主题\_\_\_\_\_基于 B 站数据的时间序列分析\_\_\_\_\_

2023 年 5 月

# 目录

1 选题背景.....	3
1.1 逐年亏损.....	3
1.2 UP 主流失 .....	4
2 选题目的.....	4
2.1 针对 B 站.....	4
2.2 针对 UP 主 .....	4
2.3 针对视频.....	5
3 时间序列分析过程.....	5
3.1 处理流程.....	5
3.2 平稳性检验.....	6
3.3 白噪声检验.....	6
3.4 ARIMA 模型及参数确定 .....	7
3.4.1 AR 模型 .....	7
3.4.2 MA 模型 .....	7
3.4.3 ARIMA 模型 .....	7
3.4.4 参数确定.....	7
3.5 模型检验.....	8
3.5.1 Q-Q 图检验 .....	8
3.5.2 Ljung-Box 检验 .....	8
4 问题解决.....	8
4.1 针对 B 站问题的解决.....	8
4.1.1 数据来源.....	8
4.1.2 平稳性检验.....	9
4.1.3 白噪声检验.....	10
4.1.4 模型选择.....	10
4.1.5 模型评价.....	11
4.1.6 模型预测.....	11
4.1.7 Auto.arima()的局限与模型优化 .....	12
4.2 针对 UP 主问题的解决 .....	13
4.2.1 数据来源.....	13
4.2.2 平稳性检验与白噪声检验.....	13
4.2.3 模型选择与评价.....	13
4.2.4 模型预测.....	14
4.3 针对视频问题的解决.....	15
4.3.1 数据来源.....	15
4.3.2 平稳性检验与白噪声检验.....	15
4.3.3 模型选择与评价.....	15
4.3.4 模型预测.....	16
5 课程心得.....	17
6 参考文献.....	17
7 附录.....	18

## 1 选题背景

B 站全名哔哩哔哩，作为从 2009 年 6 月创建的老牌视频网站，它拥有着包括动画、番剧、国创、音乐、舞蹈、游戏、知识、鬼畜等 15 个内容分区，其中，生活、游戏、动漫、科技是 B 站主要的内容品类。此外，经过十多年的发展，B 站已经不局限于视频的创作与分享，它还推出了包括直播、游戏、广告、电商、漫画、电竞等的主要业务。

随着 B 站涉足的领域越来越多，很多问题也暴露出来，主要包括两个方面：

### 1.1 逐年亏损

尽管 B 站涉猎领域众多，业务繁多，却始终没能逃离亏损的现实。财报显示，2022 财年 B 站总营收同比增长 13%，达 219 亿元，全年净亏损 75 亿元，同比扩大 10%。B 站在 UP 主培养激励、视频版权购买等方面投入过多，同时，不同于 YouTube 等老牌视频网站，B 站没有提供在其主页内投放广告的途径（例如爱奇艺站内视频观看前需要先看广告），因此收入很少。

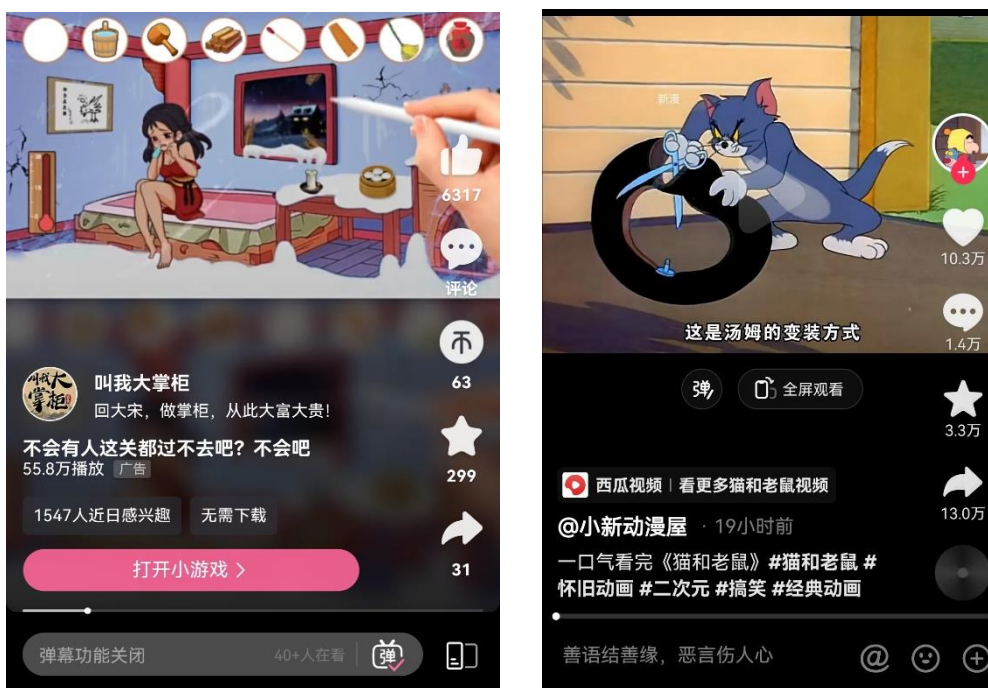


图 1 Story Mode（左 1：B 站，右 1：抖音）

为了避免常年亏损，B 站也做出了多次的转型，其中，最让用户诟病的便是 2021 年推出的“抖音化”Story Mode（短视频模式）。在模式推出后，一部分用户认为 B 站已经背离了自身“你感兴趣的视频都在 B 站”的定位，一时间，大量用户流失。

1.2 UP 主流失

作为 B 站主要构成的 UP 主，在今年的 3 月份发起了一波停更热潮，在短时间内，直接登顶微博热搜。



图 2 UP 主停更热潮

分析其原因，主要包括两个方面，第一是由于 B 站对 UP 主的创作激励逐年减少，根据“智能路障”网友的透露，两部具有相同播放量和其他数据的视频，在 2021 年得到的创作激励是 4123.18 元，而在 2023 年 1 月，就只获得了 1442.25 元；其次，短视频出现的冲击，使得长视频 UP 主的视频播放量和关注度大幅降低，很多 UP 主背后不以个人为单位，而是以公司为单位，在这种情况下，耗费大量人力、物力的视频的收入无法弥补 UP 主的开销，于是，停更热潮便出现了。



图 3 创作激励逐年减少

2 选题目的

本文基于 B 站总站数据和时间序列分析的方法，拟解决以下三个问题：

2.1 针对 B 站

根据 B 站总站的数据进行分析，可以对未来一段时间的整个 B 站的视频播放量进行分析，可以根据新推出某一转型后的一段时间的数据，对未来视频播放量数据进行预测，得到整体趋势，从而及时的对转型方式进行调整、改进，或者是暂停。

2.2 针对 UP 主

根据 UP 主的数据进行分析，得到一段时间内的粉丝数、视频播放数量等指标的变化趋势，据此预测未来一段时间内该账号是会向好的方向发展（粉丝增多

——视频播放量增多——流量增多——收入增多)还是相反。同时,根据此分析,广告商也可以找到最适合的投放广告的 UP 主,在其上升期,花费较少的投资,获得最多的引流。

### 2.3 针对视频

根据视频的播放数据、评论数、点赞率等数据进行分析,得到视频在一段时间内的热度的趋势,通过不同类型的视频进行对比,并进行时间序列分析,得到不同类型的视频在什么时间段内的播放量最高,视频的热度会在视频发布后多少天内削减到个位数播放。

## 3 时间序列分析过程

本部分将介绍基于 ARIMA 模型的时间序列分析的过程,以及部分流程中的原理,内容主要来自于文献资料和本人的学习理解。

### 3.1 处理流程

下图展示了基于统计学的时间序列模型——差分整合移动平均自回归模型(Autoregressive Integrated Moving Average)的处理流程。

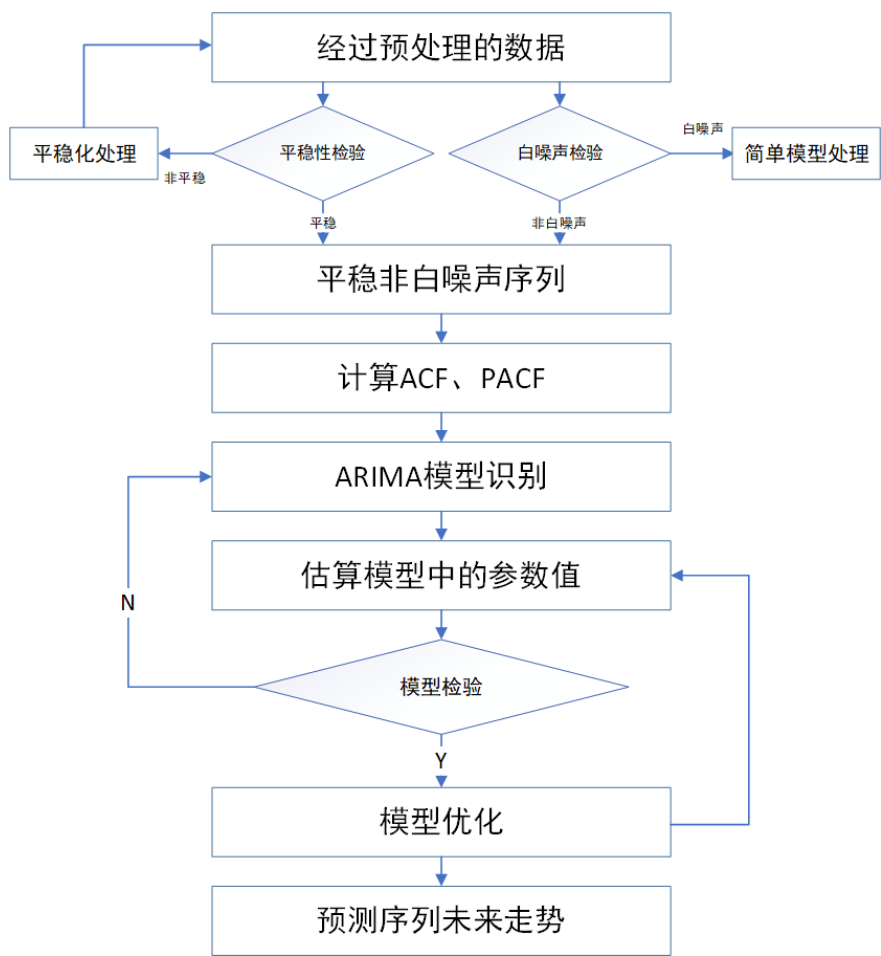


图 4 时间序列分析流程图

### 3.2 平稳性检验

大部分的时间序列统计学模型是基于时间序列是平稳性的前提。人为的主观观测数据的平稳性存在误差，所以引入了统计学中的假设检验来测试一个时间序列数据的平稳性。常见的有 ADF 检验、KPSS 检验。

本文使用的是 Augmented Dickey Fuller Test (ADF Test)，它是单位根检验(unit root test)的一种。单位根是一个使得时间序列非平稳的一个特征。

$$Y_t = \alpha Y_{t-1} + \beta X_e + \epsilon \quad (1)$$

在上面的公式中，如果  $\alpha=1$ ，那么就存在单位根。其中  $Y_t$ ， $Y_{t-1}$  分别表示  $t$  时刻和  $t-1$  时刻的时间序值， $X_e$  表示外生变量， $\epsilon$  表示误差项。

可以理解为，只有当  $\alpha < 1$  时，整个时间序列的趋势才是有可能出现逆转的。ADF test 就是对  $\alpha$  值的假设检验，它的原假设是  $\alpha = 1$ ，即原假设成立，则时间序列非平稳。

实际操作中，需观察 P 值，当 P 值小于 0.05 时，拒绝原假设，即时间序列是平稳的。当时间序列非平稳时，可以采用高阶差分和季节性调整等平稳化处理，在本文中，选用了一阶差分和高阶差分的方式。其原理大致为，将当前观测值减去前一个观测值，即将时间序列数据中的每个值减去其前一个值。这样可以得到一个新的序列，其中每个观测值表示两个时间点之间的变化，从而消除时间序列数据中的趋势和季节性成分。

### 3.3 白噪声检验

白噪声检验是用于验证时间序列中残差（或误差）是否具有随机性和独立性的统计检验方法。原假设为，时间序列的残差是白噪声序列，即具有随机性和独立性，不存在自相关或偏相关。

常见的白噪声检验有基于残差的自相关函数 (ACF) 和偏自相关函数 (PACF) 的计算的 Ljung-Box 检验或 Box-Pierce 检验。本文使用的是 Ljung-Box 检验，其步骤如下：首先，计算残差序列的 ACF 和 PACF。其次，根据给定的滞后阶数（通常为一定的滞后阶数，例如 10 或 20），计算 Ljung-Box 统计量。最后，根据计算得到的统计量和滞后阶数，进行假设检验，如果 P 值小于事先设定的显著性水平（通常为 0.05），则拒绝原假设，认为序列不是白噪声，存在自相关或偏相关。反之，序列为白噪声序列。

如果是白噪声序列，说明时间序列不包含趋势、季节性或周期性成分，因此进一步的时间序列分析可能会受到限制，可以通过增大数据量和数据质量进行改善。

### 3.4 ARIMA 模型及参数确定

ARIMA 模型是 AR 模型与 MA 模型的结合。下面先分别介绍 AR 模型、MA 模型以及两者结合的 ARMA 模型。

#### 3.4.1 AR 模型

自回归模型(Autoregressive)是拟合时间序列数据时最先尝试的模型，尤其是在对该数据没有额外的了解时。它的基本公式可以用以下公式表示，

$$y_t = \Phi_0 + \Phi_1 * y_{t-1} + \Phi_2 * y_{t-2} + \cdots + \Phi_p * y_{t-p} + e_t \quad (2)$$

这个模型称为 AR(p)，模型的最高阶为 p,即  $\Phi_p \neq 0$ ,随机干扰序列  $e_t$  为零均值的白噪声序列。

#### 3.4.2 MA 模型

移动平均模型(Moving Average)的基础是每个时刻点的值是历史数据点错误项的函数，其中这些错误项是互相独立的。MA 模型和 AR 模型的公式很类似，只是将公式中的历史数值替换成了历史数据的错误项  $e$ ，由于错误项  $e$  是互相独立的，所以在 MA 模型中， $t$  时刻的数值仅仅和最近的  $q$  个数值有关，而和更早的数据之间没有自相关性。

$$y_t = \mu + e_t + \theta_1 * e_{t-1} + \theta_2 * e_{t-2} + \cdots + \theta_q * e_{t-q} \quad (3)$$

同样的，和 AR 模型类似，满足上述公式的时间序列可以用 MA(q)来表示。

#### 3.4.3 ARIMA 模型

ARMA 模型就是 AR 和 MA 的简单结合，同时包含了历史数值项和错误项。由于 AR 模型对时间序列有平稳性要求，ARMA 模型也存在这个限制，因此我们将其拓展到 ARIMA 模型，引入的差分概念是一种获得时间序列的方法。最常用的一种差分方法是计算当前项和前项的差值，获得一组新的时间序列。

根据老师上课的讲解，在时间序列模型的选择上，可以根据 ACF 与 PACF 的变化，进行初步选择，规则如下：

表 1 时间序列模型的选取

函数	AR (p)	MA (q)	ARMA or ARIMA
ACF 曲线	缓缓下降	在间隔=q 后突然下降	没有明显截止点
PACF 曲线	在间隔=p 后突然下降	缓缓下降	没有明显截止点

#### 3.4.4 参数确定

ARIMA 有三个参数  $p$ 、 $d$ 、 $q$ ，写作 ARIMA( $p,d,q$ )，其中  $p$  代表 AR( $p$ )，自回归阶数， $d$  代表 Integrated( $d$ )，差分阶数， $q$  代表 MA( $q$ )，移动平均阶数。我们应当确保这三个参数尽可能的小避免过拟合，一个可供参数的准则是，不要让  $d$  超过 2， $p$  和  $q$  超过 5，并且  $p$  和  $q$  尽量保证一个是模型主导项，另一个相对较小。

确定参数主要包括两种方式，一种是使用自相关和偏自相关函数图（ACF 和 PACF），以及经验法则，来选择合适的滞后阶数和季节性阶数。另一种是根据已经成熟的自动选择参数的方式，进行模型拟合与定阶。前者通过 ACF 与 PACF 的图分别对 AR 模型和 MA 模型的参数进行确定，后者则通过模型拟合估计参数，计算 AIC（Akaike Information Criterion）和 BIC（Bayesian Information Criterion）的值，评价模型的好坏，最后，在一组模型中，选择最优的模型参数。

### 3.5 模型检验

在时间序列分析中，模型诊断是评估已拟合模型的残差是否符合模型假设的重要步骤。本文主要使用了两种模型检验方式，具体如下：

#### 3.5.1 Q-Q 图检验

Q-Q 图（Quantile-Quantile Plot）：通过绘制残差的 Q-Q 图，可以检查其是否符合正态分布假设。在 R 语言中，使用 `qqnorm()` 函数绘制 Q-Q 图，并使用 `qqline()` 函数添加一条直线，用于比较残差与正态分布的拟合情况。如果残差点集中在直线附近，并且基本上沿着直线分布，说明残差近似服从正态分布。

#### 3.5.2 Ljung-Box 检验

Ljung-Box 检验：通过 Ljung-Box 检验，可以检验残差序列是否存在自相关性。在 R 语言中，使用 `Box.test()` 函数进行 Ljung-Box 检验。该检验基于残差序列的自相关函数（ACF）和偏自相关函数（PACF）来评估残差序列的随机性和独立性。如果 Ljung-Box 检验的  $p$ -value 大于 0.05，则无法拒绝原假设，表明残差序列是白噪声序列，符合模型的随机性和独立性假设。

## 4 问题解决

时间序列分析解决问题的大致流程基本类似，有些过程性的步骤在解决几个问题时存在重复的现象，本文仅将不同的方法第一次出现时的问题做详细说明，后续使用同样的方法时，仅展示结果性图示。

### 4.1 针对 B 站问题的解决

对于 B 站面临的问题，本文选取了 90 天内的 B 站每日整体的视频播放量和 30 天发布的视频数的数据，来分析 B 站在一个月内热度的变化，并对未来一段时期的数据进行预测，以查看按照目前的运营政策，B 站的热度是趋于更高还是恶化。

#### 4.1.1 数据来源

本部分使用的数据来自飞瓜数据平台，平台提供的 B 站流量大盘数据主要包括 B 站整体每天的发布的视频数、播放量、点赞数与评论数。本文选取了 2023.1.23~2023.4.23 时期内 90 天的数据进行分析。



#### 4.1.2 平稳性检验

在进行平稳性检验前，我们可以通过数据可视化的方式，对数据隐藏的信息和趋势进行观测分析。

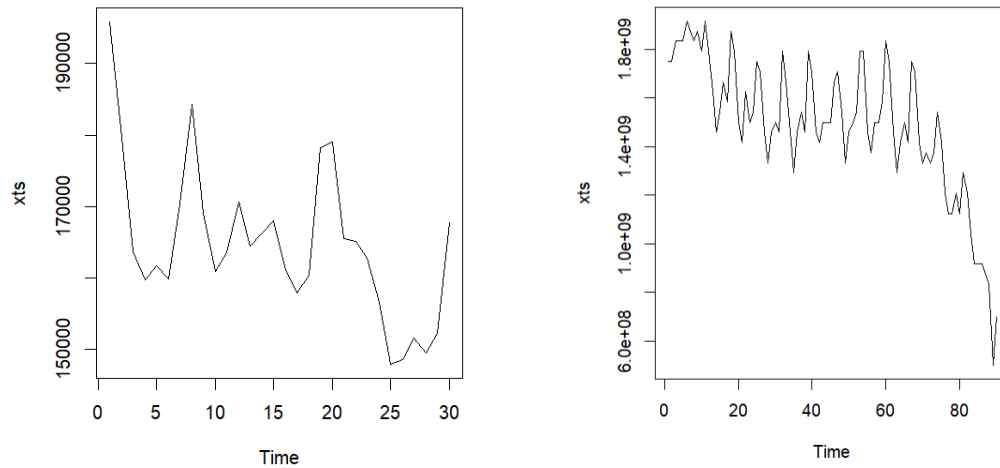


图 5 数据的可视化分析（左：视频数 右：播放量）

我们先对播放量时间序列进行平稳性检验，使用 ADF 方法，得到的结果为：

```
> xts<-ts(x1,start=1,end = 30,frequency = 1)
> ADF<-adf.test(xts)
> ADF
```

#### Augmented Dickey-Fuller Test

```
data: xts
Dickey-Fuller = -3.4325, Lag order = 3, p-value
= 0.07142
alternative hypothesis: stationary
```

图 6 ADF 平稳性检验

根据 P 值（大于 0.05），可以得出，时间序列不平稳，本文采用一阶差分的方式对时间序列进行平稳化处理，发现时间序列仍不平稳，于是采用二阶差分。

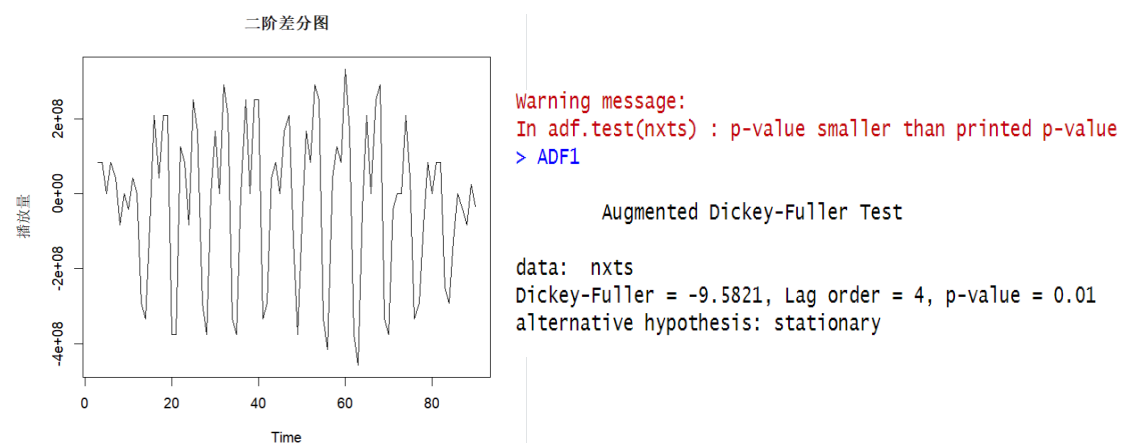


图 7 二阶差分后的时序图与平稳性检验

本文使用二阶差分对数据进行平稳化处理，再次进行 ADF 检验，此时，发现 P 值远小于 0.05，即拒绝原假设，时间序列平稳。

对于视频数数据，采用一阶差分后，数据就已经平稳了，本文不再展示。

#### 4.1.3 白噪声检验

本文使用 Ljung-Box 方法对播放量序列进行白噪声检验，得到结果如下：

```
> Box.test(nxts,type = "Ljung-Box")
```

```
Box-Ljung test

data:  nxts
X-squared = 8.5365, df = 1, p-value = 0.003481
```

图 8 白噪声检验

分析发现，P 值远小于 0.05，因此拒绝原假设，时间序列是非白噪声序列。

而在对视频数时间序列进行分析时，则发现其检验的 P 值远大于 0.05，序列是白噪声序列，因此不能继续进行检验。而出现的问题的原因也很简单，是因为数据量太少（仅有 30 条），对数据的分析造成了影响，在后续尝试中，发现将数据加大到 60 时，再进行检验，检验结果为非白噪声序列了。

此外，本文中的视频量数据也是因为原先由于数据量太少，导致后续分析无法进行，才扩充到 90 条。

#### 4.1.4 模型选择

本部分没有使用观察 ACF 与 PACF 的变化图来确定 p 和 q 值，而是直接使用了 auto.arima()，进行模型的拟合。

```
> fit<-auto.arima(xts)
> fit
Series: xts
ARIMA(0,1,0)

sigma^2 = 2.142e+16: log likelihood = -1799.63
AIC=3601.27 AICc=3601.31 BIC=3603.76
> arima<-auto.arima(xts,trace=T)

ARIMA(2,1,2) with drift          : Inf
ARIMA(0,1,0) with drift          : 3602.932
ARIMA(1,1,0) with drift          : Inf
ARIMA(0,1,1) with drift          : Inf
ARIMA(0,1,0)                    : 3601.313
ARIMA(1,1,1) with drift          : Inf

Best model: ARIMA(0,1,0)
```

图 9 模型选择

根据 AIC 值进行评价，发现结果最好的模型为 ARIMA(0,1,0)，因此选择其作为时间序列模型进行预测。

#### 4.1.5 模型评价

本文选择使用 Q-Q 图与白噪声检验进行评价，具体结果如下：

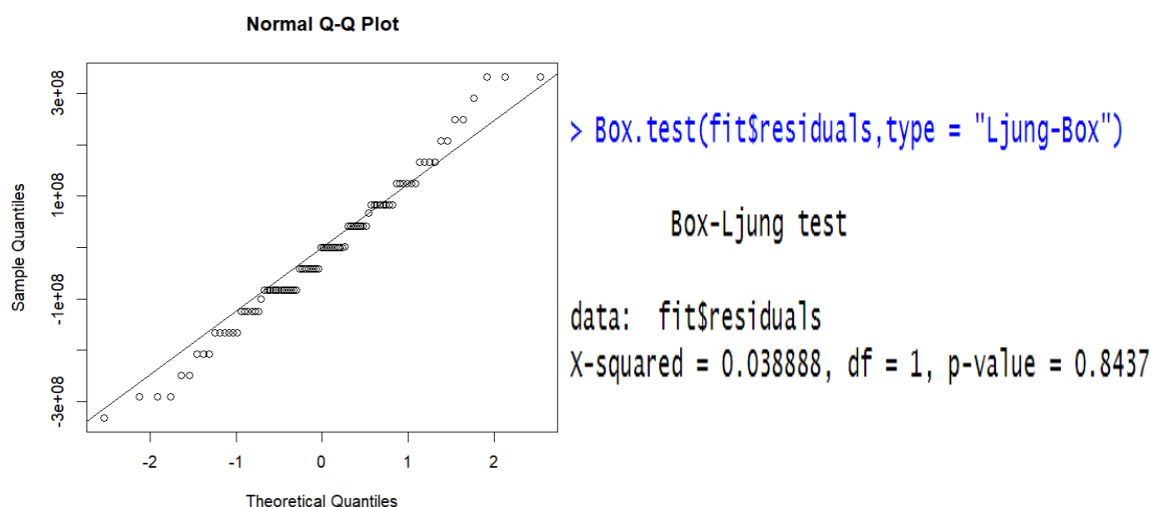


图 10 Q-Q 图检验与白噪声序列检验

根据 Q-Q 图可以看出，残差点集中在直线附近，并且基本上沿着直线分布，说明残差近似服从正态分布。说明模型 ARIMA(0,1,0)是有效的拟合模型。此外，在白噪声检测中，P 值为 0.8437，远大于 0.05，因此该序列是白噪声序列，说明模型已充分提取时间序列的信息，是有效的拟合模型。

#### 4.1.6 模型预测

在选择好合适的模型后，本文对后续 5 期的数据结果进行了预测，预测结果如图所示：

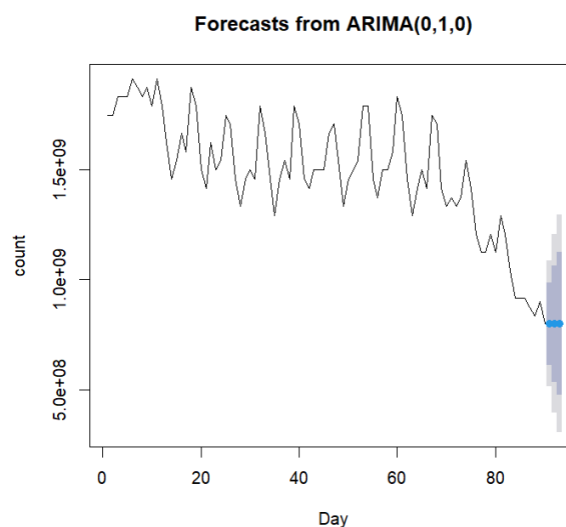


图 11 模型预测结果

根据预测结果，可以看出，B 站视频的总播放量处于下降的趋势，这说明目前 B 站的热度正在流失，B 站应调整自身的运营政策，激励 UP 发布视频，获得播放量的提升。

4.1.7 Auto.arima()的局限与模型优化

在汇报结束后，根据老师的指导，了解到了预测结果不理想（趋于直线），是因为模型的选择过于简单，说明 auto.arima 具有一定的局限性。

在优化过程中，使用了 SPSSPRO 和 Eviews 等数据分析工具，发现了比较有趣的现象。SPSSPRO 的分析结果与上文基本相同，都是选择了 ARIMA(0,1,0)模型，但相较于本文的平缓的预测结果，其结果更具有趋势性：

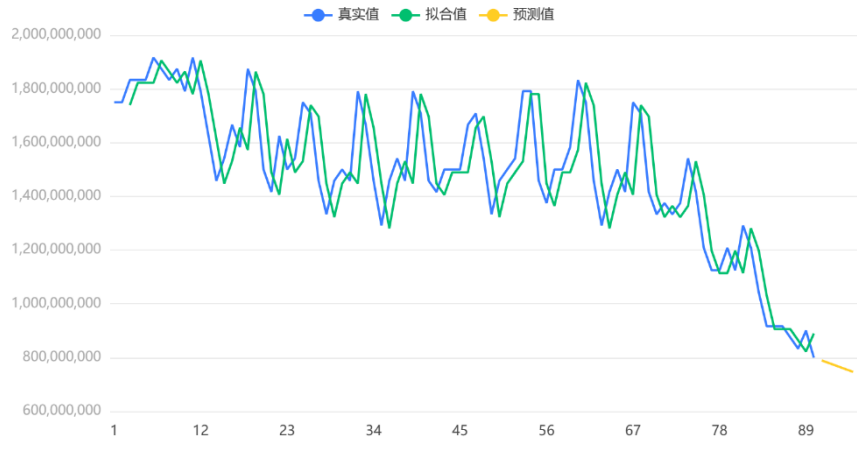


图 12 SPSSPRO 预测结果

而在 Eviews 进行预测时，从平稳性处理一步就出现了不同。由于本文选取的数据是 B 站的平台数据，因此存在一定的周期性（一周 7 天），因此，在进行差分时，选取了步长为 7 的一阶差分，在进行 ADF 检验时，Eviews 分别对带截距和趋势项、带截距、带截距和趋势项的单位根进行检验，三种单位根检验都拒绝原假设，所以判断一阶差分序列是平稳时间序列，而在 R 中使用步长为 7 的一阶差分后，ADF 检验显示序列不平稳。

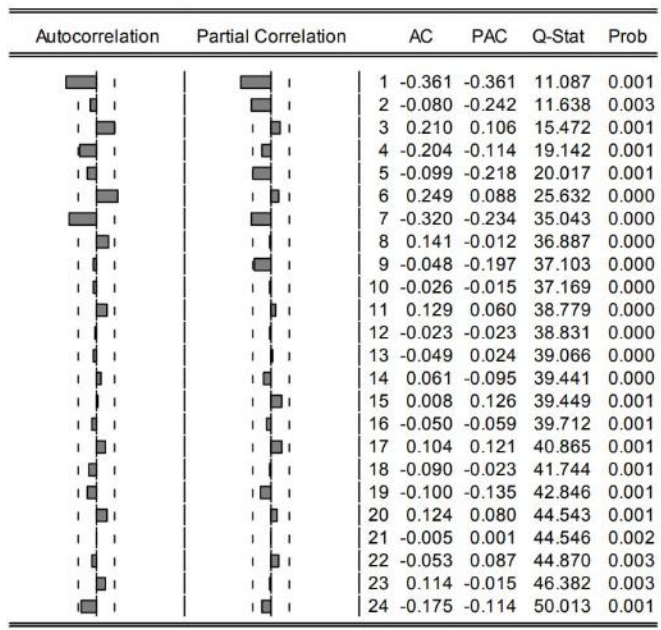


图 13 ACF 与 PACF 图

ACF 和 PACF 两者都没有明显的截尾特征，因此满足 ARIMA 建模的定义条件。根据图象，可以看出  $p$  和  $q$  的值为 1 或 2，因此考虑建立 ARIMA(1,1,1)、ARIMA(1,1,2)、ARIMA(2,1,1)、ARIMA(2,1,2)，通过 AIC 值的比较，确定了最优模型为 ARIMA(2,1,2)。经检验发现模型有效，下面是预测结果展示：

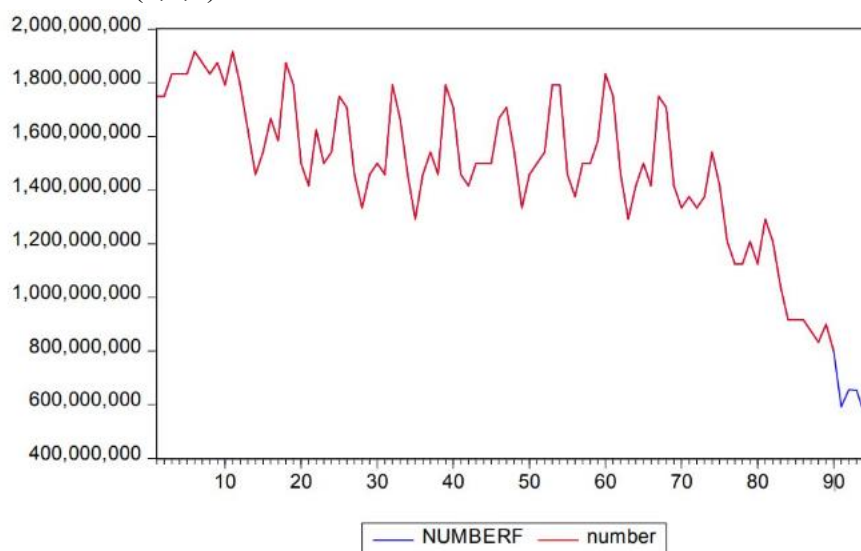


图 14 Eviews 预测结果

可以看出，使用该模型进行预测的效果是以上三种预测结果中最优的。完整的分析过程见附件。

## 4.2 针对 UP 主问题的解决

### 4.2.1 数据来源

关于 UP 主的数据，本文选取的是由番茄数据平台提供的 UP 主观测数据，其数据主要包括 UP 主一段时间内的每天的视频播放量、粉丝数变化等。本文选取了“王师傅和小毛毛”UP 主 2、3、4 月份，共 90 天的粉丝数变化数据。

### 4.2.2 平稳性检验与白噪声检验

在平稳性检验中，本部分数据的假设检验得到的  $P$  值远小于 0.05，因此符合平稳性；白噪声检验中，假设检验  $P$  值远小于 0.05，拒绝原假设，序列不是白噪声序列。

### 4.2.3 模型选择与评价

本部分使用模型拟合的方式确定 ARIMA 模型的参数，在众多模型模型 ARIMA(0,1,5)模型的 AIC 值最小，因此时间序列模型选择 ARIMA(0,1,5)。

在模型评价中，本文选择 Q-Q 图进行评价，结果如下：

残差点集中在直线附近，并且基本上沿着直线分布，说明残差近似服从正态分布。

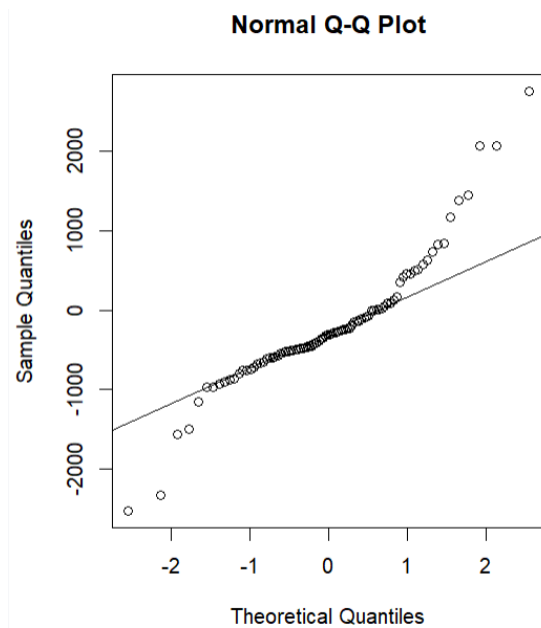


图 15 Q-Q 图模型检验

#### 4.2.4 模型预测

根据选定的模型,本文对接下来五天的该 UP 主的粉丝数变化情况进行预测,结果如下:

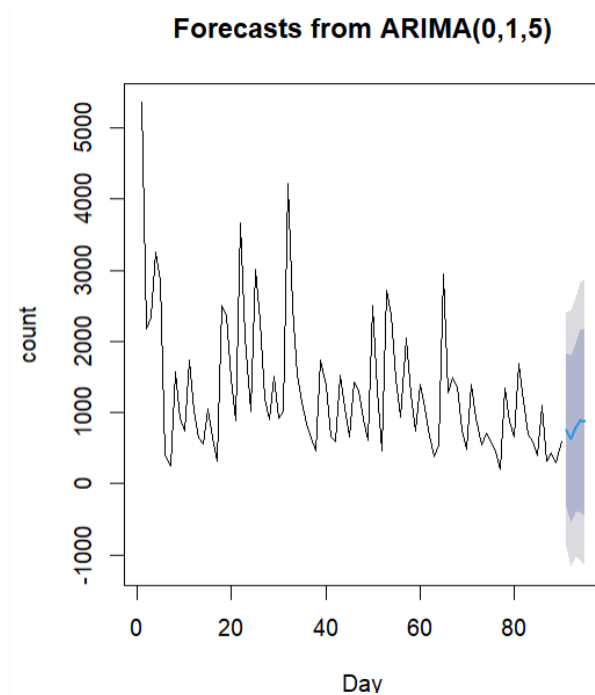


图 16 模型预测

发现在目前的发展趋势下,该 UP 主的粉丝增量数据趋于回升,这说明应该保持现阶段的发布视频策略和内容,以保证粉丝增量的回升。反之,如果粉丝增量下降,那么应寻找增加时期的视频主题与类型,做好整改工作。

### 4.3 针对视频问题的解决

#### 4.3.1 数据来源

本部分数据来自番茄数据平台的数据观测——视频检测中，本文选取了“美食区”、“科普区”两种类型的视频数据进行分析。每种视频选取其自发布以来，每 10 分钟进行一次播放量测度的数据，共选取了 1000 次测度的数据（七天）。

#### 4.3.2 平稳性检验与白噪声检验

在平稳性检验与白噪声检验中，两种类型的视频数据都在平稳化处理后通过检验，是平稳的非白噪声序列。

```
> ADF
Augmented Dickey-Fuller Test
data: xts
Dickey-Fuller = -4.5298, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary

> Box.test(xts,type = "Ljung-Box")
Box-Ljung test
data: xts
X-squared = 904.03, df = 1, p-value < 2.2e-16

> ADF
Augmented Dickey-Fuller Test
data: xts
Dickey-Fuller = -5.7986, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary

> Box.test(xts,type = "Ljung-Box")
Box-Ljung test
data: xts
X-squared = 700.76, df = 1, p-value < 2.2e-16
```

图 17 平稳性与白噪声检验

#### 4.3.3 模型选择与评价

本部分通过 Auto.arima()方法进行模型拟合，分别得到两种视频类型数据适用的 ARIMA 模型为：ARIMA(2,1,1)与模型 ARIMA(0,1,2)。通过使用 Q-Q 图与白噪声检验的方式，发现模型拟合效果良好，模型有效。

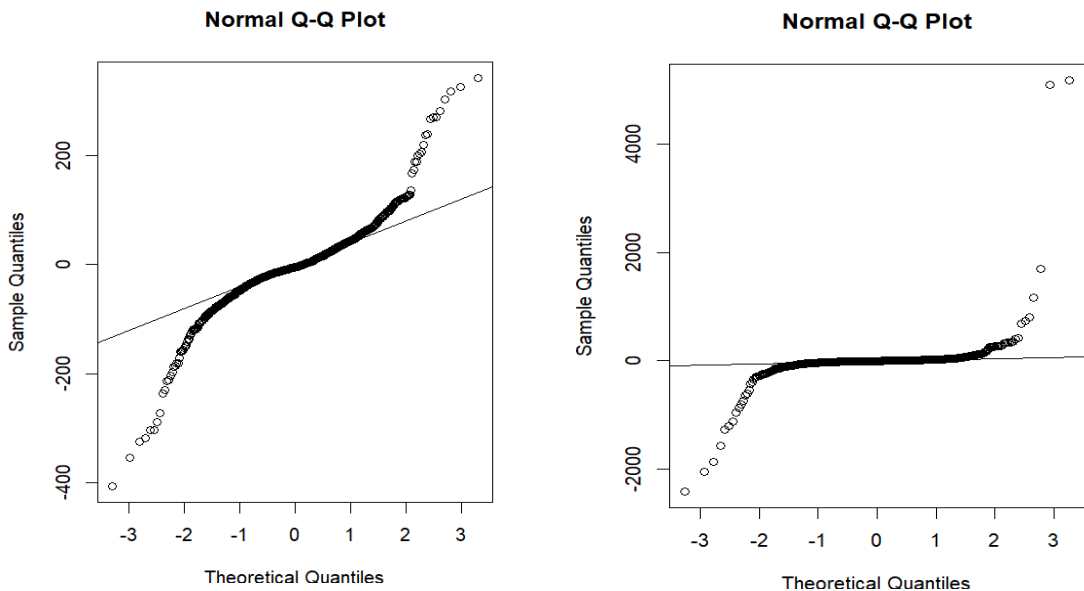


图 18 Q-Q 图检验

#### 4.3.4 模型预测

下面是使用模型对两种视频类型数据进行往后 20 期、30 期预测的结果：

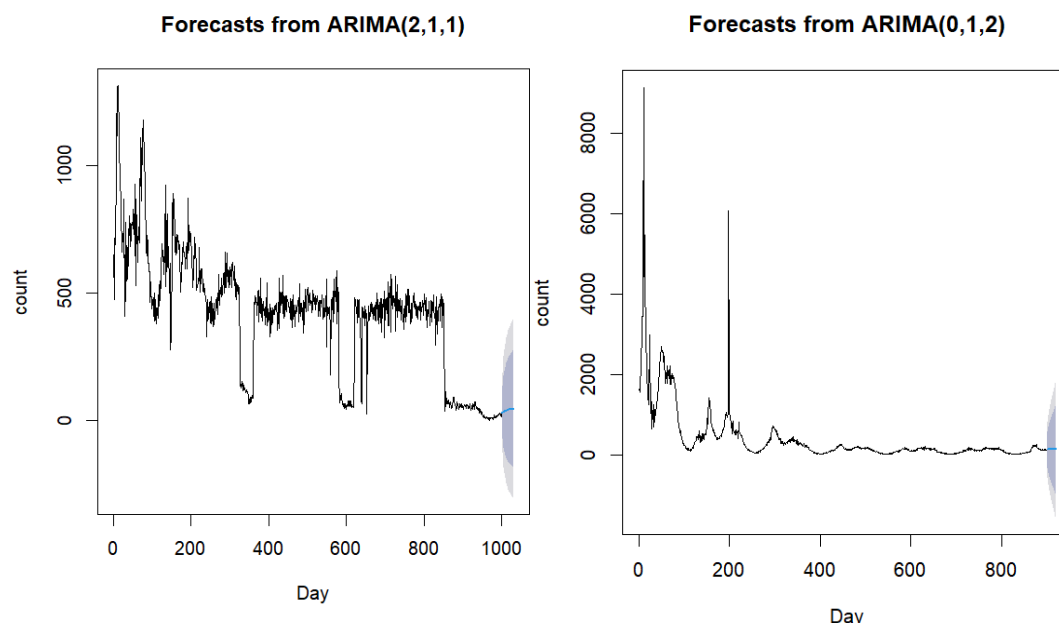


图 19 预测结果展示

左边的预测结果是科普类视频的播放量，可以看出该类视频的播放量在除了夜间时间段内持续很高且具有很强的规律性，尽管随着发布时间变化逐渐降低，但也保持着较高的水平，未来 300 分钟内（每期 10 分钟共 30 期）播放量处于上升趋势。

右边预测结果是美食区视频的播放量，可以看出，该类视频的播放量在刚播出的时候，存在非常高的热度，在发布一段时间内，热度逐渐下降，每天的时候都存在一段时间的高峰期，主要集中在“饭点”。整体的播放量数据和科普类视频相比较差，在预测中，也可以发现，视频的播放量趋于下降。

通过对比，可以发现，不同类型的视频类型在播放量数据上的差距较大，而且不同类型的视频的用户的观看时间不相同，比如，美食区的视频大家可能都是在吃饭的时候观看，因此播放量最高的时间集中在“饭点”，对应的，美食区 UP 主可以在“饭点”进行发布视频，以获得最多的播放量。美食区视频的第二个播放量高峰期出现的原因是，在视频发布的第二天，UP 对自己前一天发布的视频进行了转发，因此视频又在该 UP 主的动态中显示，因此更多的人关注到视频。对于科普区视频，可以发现，在视频发布的第 6 天左右时，视频的热度降低很多，可以通过类似转发视频的方式，增加视频热度。



## 5 课程心得

最早接触到时间序列分析的概念，是在大二的选修课《预测与决策》中，当时由邓老师为我们讲解了基本的 AR、MA、ARIMA 模型和根据 ACF 与 PACF 图进行参数的确定，以及一阶差分和二阶差分的计算。当时没有使用软件进行计算，都是选择简单的模型，然后进行人工判断与计算。同时，在模型的原理方面，由于还没有怎么接触统计学，所以当时上课的时候很痛苦，原理方面基本不懂，只能根据 ACF 的图大致确定一下参数，进行简单的应用。

在这学期的《时间序列分析》课程中，韦老师重新细致的为我们讲解了时间序列的原理和使用，在每一部分结束后，都有对应的 R 语言程序的上机实验，通过代码的实践，我对不同数据的特征有了大致的了解，比如可以大致根据时序图或者一阶差分后的时序图判断序列是否平稳等。同时，对是时间序列分析方法的整个的发展过程有了大致的了解，包括统计学方面的时间序列分析、基于机器学习与深度学习的时间序列分析等，对时间序列的整体分析流程有了深入理解，对每一步的原理也有了自己的理解，可以说是收货颇多。

在完成大作业的过程中，也遇到了很多的问题，本来准备借助 Chatgpt 给出的代码和 CSDN 上给出的完整代码示例稍作修改完成作业。但是在实际操作的过程中，才发现很多问题需要具体分析。比如 `auto.arima` 模型具有一定的局限性，需要人工根据 ACF 数据进行参数调整，比如在第一部分的预测中，对视频数进行预测时，在验证白噪声序列时，发现序列就是白噪声序列，没法继续完成，只能查找原因，更换数据；在进行预测时，发现由于模型选择简单（ARIMA(0,1,0) 模型），导致模型预测是一条水平线，需要对模型进行优化；在优化过程中，发现不同应用给出的一阶差分结果的平稳化检验不相同等问题，这些都需要根据情况进行改进。所以在完成作业的过程中，由于给的时间比较长，所以不断完善，一写便停不下来，导致了篇幅越来越多，这也是自上学期多元统计分析后，我写过的最认真的课程报告。总之，在完成课程报告的过程中，对时间序列分析过程有了深刻的理解，并且收货很多。





## 6 参考文献

- [1] 汤岩. 时间序列分析的研究与应用[D].东北农业大学,2007.
- [2] 赵国顺. 基于时间序列分析的股票价格趋势预测研究[D].厦门大学,2009.
- [3] 罗芳琼,吴春梅.时间序列分析的理论与应用综述[J].柳州师专学报,2009,24(03):113-117.
- [4] [时间序列分析 101: 序言 - TimeSeriesAnalysis101 \(gitbook.io\)](https://time-series-analysis-101.gitbook.io/)

## 7 附录

### 7.1 数据

在“数据”文件夹中，保存了本文进行时间序列分析使用的数据，具体包括：

名称	修改日期	类型	大小
 B站总站数据.xlsx	2023/5/22 20:25	Microsoft Excel 工作表	11 KB
 UP主粉丝数变化.xlsx	2023/5/22 20:42	Microsoft Excel 工作表	10 KB
 视频1.xlsx	2023/5/22 21:17	Microsoft Excel 工作表	71 KB
 视频2.xlsx	2023/5/22 21:27	Microsoft Excel 工作表	68 KB

其中，从上往下，第一个是 B 站的视频播放量数据，第二个是 UP 主粉丝变化的观测数据，视频 1 是科普类视频数据，视频 2 是美食类视频数据。



### 7.2 代码

在“代码”文件夹中，存放着本文进行时间序列分析使用的 R 语言代码，具体代码作用已以命名的方式给出，具体包括：

名称	修改日期	类型	大小
 时间序列分析 (视频播放量) .R	2023/5/22 21:11	R 文件	2 KB
 时间序列分析 (视频问题解决 (视频1) ...	2023/5/23 14:13	R 文件	1 KB
 时间序列分析 (视频问题解决 (视频2) ...	2023/5/23 14:12	R 文件	1 KB
 时间序列预测 (UP主问题的解决) .R	2023/5/22 21:08	R 文件	1 KB

### 7.3 模型优化

在“模型优化”文件夹中，存放着本文在进行对 B 站视频播放量变化进行时间序列分析时的使用不同平台进行模型优化的结果，具体如下：

名称	修改日期	类型	大小
 B站总站数据的ARIMA模型分析 (Eviews) .pdf	2023/5/23 16:39	Microsoft Edge ...	182 KB
 B站总站数据的ARIMA模型分析 (SPSSPRO) .pdf	2023/5/23 15:29	Microsoft Edge ...	5,016 KB

其中，SPSSPRO 平台文件可以通过直接导出得到，Eviews 得出的结果由本人分析整理后编写得到。

### 7.4 汇报 PPT

在和报告的同级目录内，存放着本次大作业的汇报 PPT。