



東北大學 秦皇島分校  
Northeastern University at Qinhuangdao

## 概率论与数理统计 (含随机过程) 结课论文

新冠疫情对秦皇岛旅游业影响的统计测度研究

院 别	数 学 与 统 计 学 院
专业名称	数 据 科 学 与 大 数 据 技 术
班级学号	202015140
学生姓名	周 华
指导教师	张 贵 来

2022 年 6 月 15 日



## 新冠疫情对秦皇岛旅游业影响的统计测度研究

### 摘 要

2020 年爆发的新冠肺炎疫情对秦皇岛乃至全国的旅游业造成了巨大冲击。秦皇岛是著名的滨海旅游、休闲、度假胜地,旅游资源丰富。秦皇岛旅游业是秦皇岛的支柱产业之一,拉动了第三产业的发展。研究新冠肺炎疫情对秦皇岛旅游业的影响对未来秦皇岛旅游业及服务业发展、减小损失具有重要意义。

**第一步:** 利用 2012 年 1 月至 2019 年 12 月的秦皇岛接待游客总量数据建立 SARIMA 模型,然后根据模型得出疫情爆发后 2020 年 1 月至 2021 年 12 月接待游客总量的预测值,根据实际值与预测值的差距得到影响结果:**2020 年秦皇岛旅游接待总人数减少 80.87%,2021 年秦皇岛旅游接待总人数减少 78.91%,两年平均减少 79.89%。**

**第二步:** 使用 2012 年 1 月-2019 年 12 月的数据作为训练集训练 LSTM 神经网络,每 24 个月的数据用于预测下一个月的数据,然后使用 2020,2021 年的数据作为输入,即疫情发生后 24 个月的月作为输入,逐步预测 2022 年的数据,最终与疫情发生前预测的 2022 年数据,以及实际的 2020,2021 年做对比,得出结论:**预计疫情对秦皇岛 2022 年旅游业的冲击依然很大,相比疫情前减少 65.47%;相比 2020 与 2021 年,疫情对秦皇岛的旅游业影响变小,秦皇岛旅游业相比 2020,2021 呈上升趋势,预计比 2021 年增长 31.76%。**

**关键词:** 新冠疫情;接待游客;SARIMA 模型;LSTM 神经网络;影响因子



## 目 录

1 研究背景与思路	1
1.1 研究背景	1
1.2 研究思路	1
2 SARIMA 模型和神经网络回归模型	2
2.1 SARIMA 模型	2
2.1.1 自回归模型 ( $AR(p)$ )	2
2.1.2 滑动和模型 ( $MA(q)$ )	3
2.1.3 自回归滑动和模型 ( $ARMA(p,q)$ )	3
2.1.4 ARIMA 模型	3
2.1.5 SARIMA 模型	4
2.2 神经网络回归模型	4
2.3 神经网络	4
2.4 LSTM	5
3 秦皇岛旅游接待人数的 SARIMA 模型	7
3.1 指标选取及数据预处理	7
3.1.1 指标选取	7
3.1.2 数据预处理	8
3.2 时间序列预处理	8
3.2.1 平稳性检验	8
3.2.2 白噪声检验	10
3.3 模型构建	11
3.3.1 初选模型	11
3.3.2 参数估计	11
3.3.3 残差分析	11
4 基于神经网络的秦皇岛旅游接待人数的影响和预测回归模型	12
4.1 2020、2021 疫情对秦皇岛旅游业接待人数的影响程度	12
4.2 预测 2022 年疫情对秦皇岛旅游业接待人数影响程度	14
4.2.1 基于 lstm 神经网络构建回归预测模型	14



## 概率论与数理统计(含随机过程) 结课论文

---

4.2.2 基于回归模型对 2022 年影响程度进行预测 .....	15
5 主要结论和存在问题 .....	16
5.1 主要结论 .....	16
5.2 存在问题 .....	16
参考文献 .....	17



## 1 研究背景与思路

### 1.1 研究背景

新型冠状病毒肺炎 (Corona Virus Disease 2019, COVID-19), 简称“新冠肺炎”, 世界卫生组织命名为“2019 冠状病毒病”, 是指 2019 新型冠状病毒感染导致的肺炎。2019 年 12 月截至欧洲中部夏令时间 2022 年 6 月 8 日, 全球累计新冠肺炎确诊病例 5.3 亿例, 累计死亡病例 630 万例。

秦皇岛, 是河北省辖地级市, 国务院批复确定的中国环渤海地区重要的港口城市, 著名的滨海旅游、休闲、度假胜地, 因秦始皇东巡至此派人入海求仙而得名, 是中国唯一一个因皇帝帝号而得名的城市, 因《浪淘沙·北戴河》而闻名遐迩, 汇集了丰富的旅游资源, 是驰名中外的旅游休闲胜地, 山海关区是国家历史文化名城。

秦皇岛旅游业是秦皇岛的支柱产业之一, 拉动了第三产业的发展, 占据了秦皇岛 GDP 的相当比重。新冠疫情对人民群众生命健康安全造成了严重威胁, 经济停滞, 物流受阻, 对秦皇岛, 中国乃至世界旅游业造成了巨大冲击。从长期看, 冲击是暂时的, 不会对我国旅游产业的基本面造成改变。中国共产党始终把人民群众生命安全和身体健康放在第一位, 在党中央高度重视、科学部署下, 新冠疫情得到有效控制。2020 年中国 GDP 为 101.4 万亿元, 比上年增长 2.2%, 是全球主要经济体中唯一实现经济正增长的国家。2021 年中国 GDP 总量达到 114.4 万亿元人民币, 实际 GDP 同比增长 8.1%。

国内疫情控制局面良好的背景下, 秦皇岛旅游业也逐渐回暖。但在 2022 年, 新冠病毒变异株 Omicron 迅速席卷全球, 输入到中国境内。2022 年 2 月秦皇岛毗邻城市葫芦岛疫情, 3 月长春疫情, 4 月上海疫情再次对秦皇岛乃至中国旅游业造成冲击。疫情对 2020, 2021 年秦皇岛旅游业的影响有多大? 影响的趋势如何? 因此, 研究新冠疫情对秦皇岛旅游业的影响具有重要意义。

### 1.2 研究思路

本文通过 2012-2019 年秦皇岛接待游客总数构建 SARIMA 模型, 使用 SARIMA 模型对 2020, 2021 年接待游客总数, 通过与实际接待游客总数进行对比, 探究新冠疫情对秦皇岛 2020, 2021 年旅游业的影响。再通过 SARIMA 模型得出的新冠疫情对 2020, 2021 秦皇岛影响的数据确定两年每个月的影响因子。通过对 24 个影响因子进行神经网络回归, 进而预测 2022 年每个月的影响因子, 从而预测 2022 年旅游业的发展趋势。



具体研究思路如下：

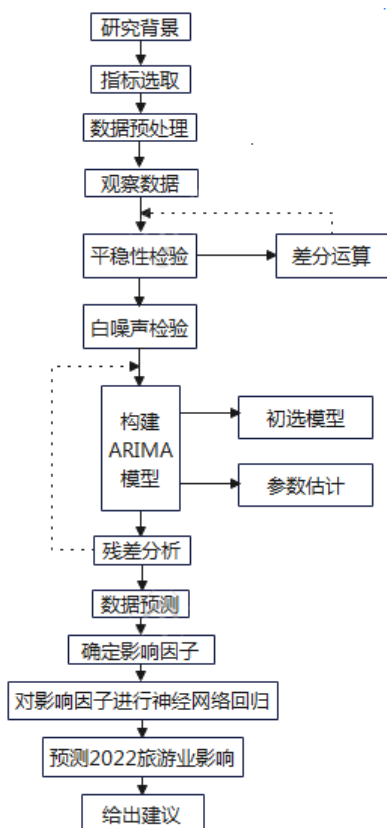


图 1.1 论文研究思路

## 2 SARIMA 模型和神经网络回归模型

### 2.1 SARIMA 模型

#### 2.1.1 自回归模型 (AR(p))

对于一个时间序列  $\{\xi_t : t \in N\}$ ，如果对任意  $t \in N, \{\xi_t\}$  满足

$$1^\circ \quad \xi_t = \varphi_1 \xi_{t-1} + \cdots + \varphi_p \xi_{t-p} + \varepsilon_t = \sum_{i=1}^p \varphi_i \xi_{t-i} + \varepsilon_t \quad (2.1)$$

$$2^\circ \quad \forall t < s, E(\xi_t \varepsilon_s) = 0 \quad (2.2)$$

则称  $\{\xi_t\}$  满足自回归模型，其中  $p > 0, \varphi_p \neq 0, \{\varepsilon_t\}$  为白噪声，即  $E(\varepsilon_t \cdot \varepsilon_s) = 0 (t \neq s), E\varepsilon_t^2 = \sigma_\varepsilon^2 > 0, \varphi_i (i = 1, 2, \cdots, p)$  为常数； $p$  称为自回归模型的阶数。自回归模型用  $AR(p)$  表示。

不难看出  $\{\xi_t\}$  是当前的一个信号，它与前面  $p$  个信号有关，而  $\{\varepsilon_t\}$  是当前信号的干扰，它不会干扰到当前信号之前的信号。

### 2.1.2 滑动和模型 ((MA(q))

如果时间序列  $\{\xi_i\}$  对任一  $t \in \mathbf{N}$ , 有

$$1^\circ \quad \xi_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} = - \sum_{i=0}^q \theta_i \varepsilon_{t-i} \quad (2.3)$$

$$2^\circ \quad \forall t < s, E(\xi_t \cdot \varepsilon_s) = 0. \quad (2.4)$$

其中,  $(\theta_0 = -1, \theta_i (i = 1, 2, \cdots, q)$  为常数,  $\theta_q \neq 0$ ); 称  $\{\xi_t\}$  满足滑动和模型, 用 MA (q) 表示上述滑动和模型, q 称为 MA(q) 的阶数.

### 2.1.3 自回归滑动和模型 ((ARMA(p,q))

如果时间序列  $\{\xi_i\}$  对任一  $t \in \mathbf{N}$ , 有

$$1^\circ \quad \xi_t - \varphi_1 \xi_{t-1} - \cdots - \varphi_p \xi_{t-p} = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}, \quad (2.5)$$

$$2^\circ \quad \forall t < s, E(\xi_t \cdot \varepsilon_s) = 0. \quad (2.6)$$

其中  $\varphi_1, \varphi_2, \cdots, \varphi_p, \theta_1, \theta_2, \cdots, \theta_q$  均为常数;  $\varphi_p \neq 0, \theta_q \neq 0$ ; 则  $\{\xi_t\}$  称为自回归滑动和序列, 或称  $\{\xi_t\}$  满足自回归滑动和模型. 记为 ARMA(p, q), (p, q) 称为模型的阶数. 不难看出 ARMA(p, q) 模型包含了 AR(p) 和 MA(q)。

### 2.1.4 ARIMA 模型

ARIMA 模型在 ARMA 的基础上引入了差分运算。首先介绍差分运算。<sup>[6]</sup>

相距一期的两个序列使之间的减法运算称为 1 阶差分运算。记  $\nabla x_t$  为  $x_t$  的 1 阶差分:

$$\nabla x_t = x_t - x_{t-1}$$

对 1 阶差分后序列再进行一次 1 阶差分运算称为 2 阶差分。记  $\nabla^2 x_t$  为  $x_t$  的 2 阶差分:

$$\nabla^2 x_t = \nabla x_t - \nabla x_{t-1}$$

依此类推, 对 p-1 阶差分后序列再进行一次 1 阶差分运算称为 p 阶差分。记  $\nabla^p x_t$  为  $x_t$  的 p 阶差分:

$$\nabla^p x_t = \nabla^{p-1} x_t - \nabla^{p-1} x_{t-1}$$

假设 p, q, d 已知, ARIMA(p, d, q) 可以表示为:

$$\nabla^d y_t = \varphi_0 + \varphi_1 \nabla^d y_{t-1} + \cdots + \varphi_p \nabla^d y_{t-p} + \varepsilon - \theta_1 \varepsilon_{t-1} - \cdots + \theta_q \varepsilon_{t-q}$$

其中,  $\varphi$  表示 AR 的系数,  $\varepsilon$  表示 MA 的系数。

## 2.1.5 SARIMA 模型

SARIMA 模型是在 ARIMA 模型上的进一步扩展, SARIMA 通过周期间隔上做 ARIMA, 可以适应季节性的时间序列, SARIMA 模型可以表示为:

$$ARIMA(p, d, q) * (P, D, Q)^s$$

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。

## 2.2 神经网络回归模型

## 2.3 神经网络

神经网络是一种模仿动物神经元而产生的模型, 可用于分类, 回归等问题, 模型效果良好。神经网络在理论上能够拟合任意函数, 包括非线性函数, 是当下最流行的数学建模方法之一。

1959 年两个生物科学家发现青蛙的神经元接受多个输入, 输入包括青蛙的多个器官的输入, 只有单输入的和到达一个阈值, 才会有输出 (青蛙接受的刺激比较大时才会有反应。) 于是计算机科学家仿照生物神经元的原理和结构, 提出了感知器, 感知机可以表示为

$$y = o\left(\sum_{i=0}^n w_i x_i + w_0\right),$$

其中

$$o(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

在感知机的基础上, 科学家们设置了输入层, 隐藏层, 输出层, 并在每一层设置若干

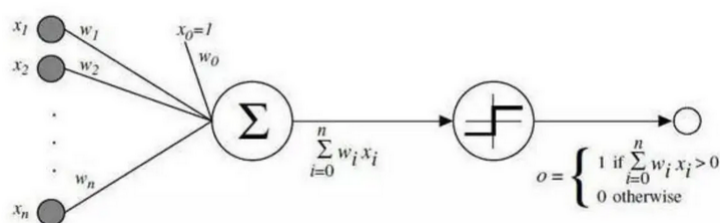


图 2.1 感知器



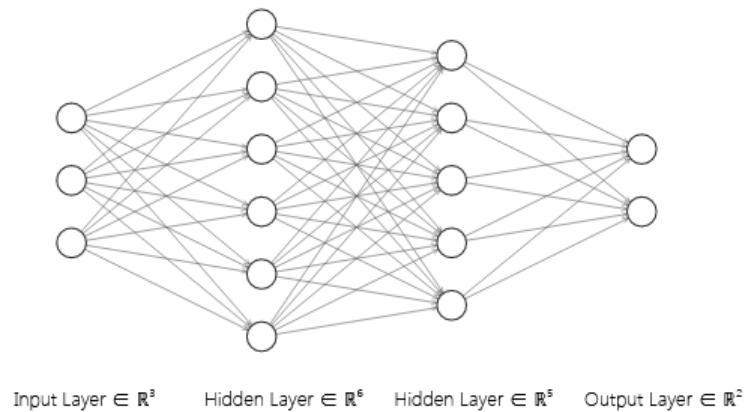


图 2.2 神经网络

个感知机，形成了神经网络。神经网络使用剃度下降法求解最优值，剃度下降法是一类求解函数最小值的计算机算法。假设希望求解目标函数  $f(x) = f(x_1, \dots, x_n)$  的最小值，可以从一个初始点  $x^{(0)} = (x_1^{(0)}, \dots, x_n^{(0)})$  开始，基于学习率  $\alpha > 0$  构建一个迭代过程：

$$\begin{aligned} x_1^{i+1} &= x_1^i + \alpha \frac{\partial f}{\partial x_1}(x^{(i)}) \\ &\vdots \\ x_n^{i+1} &= x_n^i + \alpha \frac{\partial f}{\partial x_n}(x^{(i)}) \end{aligned}$$

其中  $x^{(i)} = (x_1^{(i)}, \dots, x_n^{(i)})$ ,  $i \geq 0$ , 一旦达到收敛条件的话，迭代就结束了。图 2.3是用剃度下降法求解  $y = x^2$  的最小值时迭代 10 次的过程，其中的点代表每次迭代后的值。

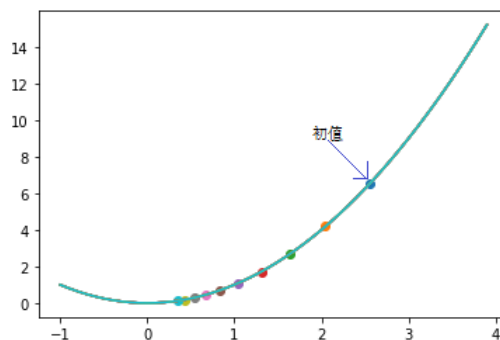


图 2.3 剃度下降法图解

## 2.4 LSTM

LSTM, 又称长短记忆单元，是在神经网络基础上发展形成的模型，可以提取长期序列的信息并有一定的遗忘功能。LSTM 主要有三个门：忘记门，输入门，输出门。

忘记门用于过滤信息，决定了从状态中丢弃什么信息：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

忘记门会读取前一序列中的输出  $h_{t-1}$  和当前模型的输入  $x_t$ ，来控制细胞状态  $C_{t-1}$  中的每个数字是否保留。其中,  $f_t$  代表忘记门的输出结果,  $\sigma$  代表激活函数,  $W_f$  代表忘记门的

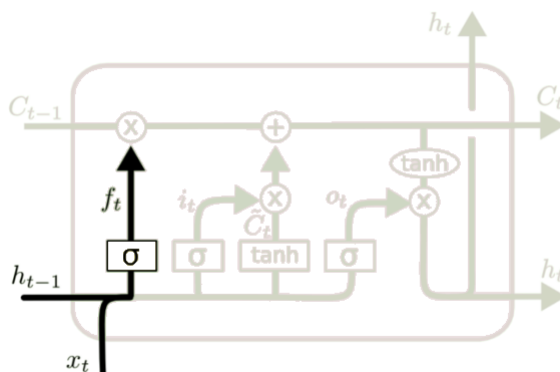


图 2.4 遗忘门结构

权重,  $x_t$  代表当前模型的输入,  $h_{t-1}$  代表前一个序列模型的输出,  $b_f$  代表忘记门的偏置。

输入门部分包括输入门和输入门更新。输入门公式为：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$C_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

输入门状态更新公式为：

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

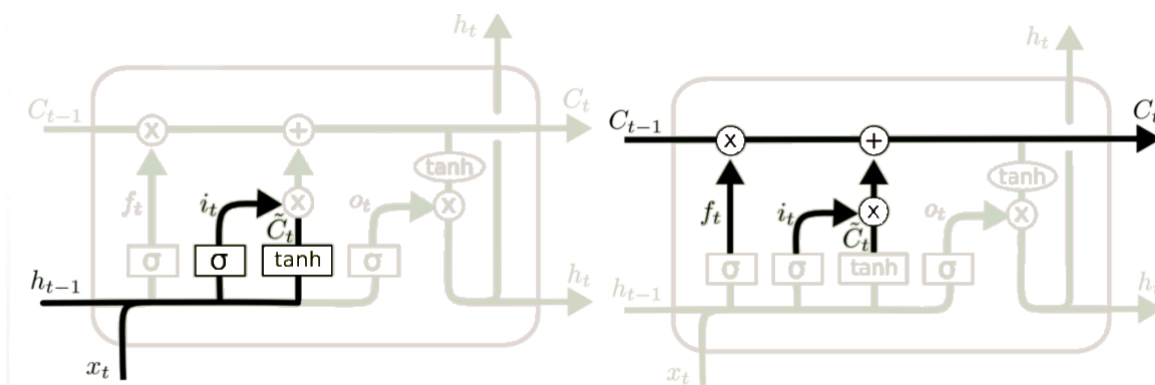


图 2.5 输入门和输入门更新

忘记门找到了需要忘掉的信息  $f_t$  后, 再将它与旧状态相乘, 丢弃确定需要丢弃的信息。然后, 将结果加上  $i_t \times \tilde{C}_t$  使细胞状态获得新的信息。这样就完成了细胞状态的更新。

输出门用于输出想要的部分, 公式如下:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (2.7)$$

$$h_t = o_t * \tanh(C_t) \quad (2.8)$$

在输出门中, 通过一个激活函数层 (实际使用的是 *Sigmoid* 激活函数) 来确定哪个部分的信息将输出, 接着把细胞状态通过  $\tanh$  进行处理 (得到一个在  $-1 \sim 1$  的值), 并将它和 *Sigmoid* 门的输出相乘, 得出最终想要输出的那个部分。

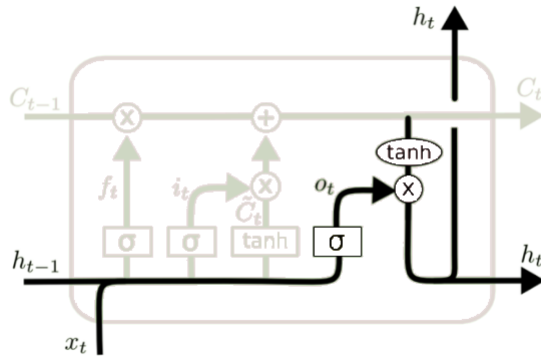


图 2.6 输出门

### 3 秦皇岛旅游接待人数的 SARIMA 模型

#### 3.1 指标选取及数据预处理

##### 3.1.1 指标选取

根据秦皇岛统计局的数据, 本文初步选取了旅游收入和接待游客数量作为衡量秦皇岛旅游数据的指标, 其中旅游收入包括国内收入和外汇收入, 接待人数包括国内游客人数和国外游客人数。最终选择接待游客总人数作为衡量指标, 即国内收入和外汇收入的和, 原因如下: (1) 旅行过程中旅客的所有消费都会记录为旅游收入, 相比旅游收入, 接待人数具有更好的准确性。(2) 秦皇岛统计局给出的旅游收入统计口径不一, 缺失的数据较多。(3) 根据历年数据, 国外游客的数量远远少于国内游客, 平均仅占 0.722%, 加之国外疫情严重, 来华游客大幅减少, 部分缺失外国游客接待数的数据, 可以将国内游客接待数近似看成总接待人数。

## 3.1.2 数据预处理

秦皇岛 2012-2021 年旅游接待人数数据具有相当一部分的缺失。本文主要采用了一下四种方法进行填充：

**(1) 逻辑推理** 如果知道  $y$  年  $m$  月接待游客总数为  $sum$ ，增长率为  $rate$ ，则  $y-1$  年  $m$  月的接待游客总数  $sum_{y-1}$  为：

$$sum_{y-1} = \frac{sum}{1 + rate}$$

**(2) 均值填充** 由于统计局的统计口径是每年 2 月开始统计累计值，如果一年中仅缺失前  $a(a \leq 3)$  个月数据的，采用平均值填充

$$sum_i = sum_{a+1} \frac{i}{a+1}, \quad i \leq a$$

**(3) 平均值填充** 不属于第二种情况，一年中只缺失一个数据并且不是最后一个月的，使用前后两个月数据进行平均值填充：

$$sum_i = \frac{sum_{i-1} + sum_{i+1}}{2}$$

**(4) 平均增长率填充** 一年缺失数据大于 3 个且不满足前面三个的，假设当年缺失数据为  $sum_1, sum_2, \dots, sum_a$ ，已知当年已知数据的增长率  $rate_{a+1}, rate_{a+2}, \dots, rate_{12-a}$  以及缺失数据上一年对应月的数据  $lastsum_1, lastsum_2, \dots, lastsum_a$ ，则

$$sum_i = lastsum_i \frac{\sum_{i=a+1}^{12-a} rate_i}{12-a}$$

## 3.2 时间序列预处理

### 3.2.1 平稳性检验

根据 2012 年-2021 年每月秦皇岛接待游客总人数数据，绘制曲线图如下：

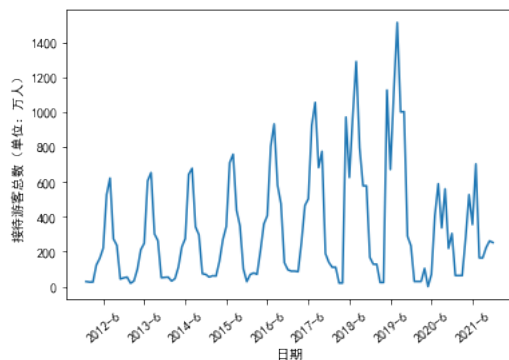


图 3.1 秦皇岛 2012-2021 接待游客总数趋势

## 概率论与数理统计(含随机过程) 结课论文

从曲线图中可以看出, 2020, 2021 年旅游接待总人数明显下降, 说明秦皇岛旅游业确实受到了疫情的冲击。根据曲线图, 明显看出改序列不是一个平稳序列, 长期具有上升的趋势, 并且受季节周期性波动, 不符合时间序列建模条件。因此对 2012-1 到 2019-12 数据进行一阶季节 12 步差分, 即  $sum_i = sum_{i+12} - sum_i$ , 得到的数据如图所示:

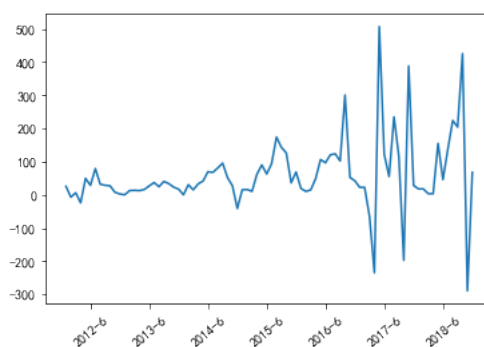


图 3.2 秦皇岛 2012-2019 接待游客总数一阶季节差分趋势图

经过一阶逐步差分和一阶季节差分, 数据已经消除了趋势性和周期性, 下面进行数据平稳性检验。假设原始数据  $\{\xi_i\}$  的  $AR(p)$  模型为:

$$\xi_t = \varphi_1 \xi_{t-1} + \cdots + \varphi_p \xi_{t-p} + \varepsilon_t,$$

则令

$$\rho = \varphi_1 + \varphi_2 + \cdots + \varphi_p - 1$$

$$H_0 : \rho < 0 (\text{序列 } \xi_t \text{ 平稳}) \leftrightarrow H_1 : \rho = 0 (\text{序列 } \xi_t \text{ 非平稳})$$

ADF 检验统计量为:

$$\tau = \frac{\hat{\rho}}{S(\hat{\rho})}$$

式中,  $S(\hat{\rho})$  为参数  $\rho$  的样本标准差。使用 EVIEW6.0 软件求得 ADF 值及临界值如下:

		t-Statistic	Prob.*
<b>Augmented Dickey-Fuller test statistic</b>		<b>-9.604360</b>	<b>0.0000</b>
<b>Test critical values:</b>	<b>1% level</b>	<b>-4.090602</b>	
	<b>5% level</b>	<b>-3.473447</b>	
	<b>10% level</b>	<b>-3.163967</b>	

图 3.3 ADF 平稳性检验

得 ADF 的检验值为  $-9.60$ , 小于 1% 的临界值  $-4.09$ , 接受原假设, 认为序列平稳。

## 3.2.2 白噪声检验

对于时间序列  $\{x_i\}$ , 任意取观察期数为  $n$  的观察序列  $\{x_t, t = 1, 2, \dots, n\}$ , 该样本的  $k$  阶非零延迟期数自相关系数为  $\hat{\rho}_k$ , 则

$H_0$ : 至少存在某个  $\rho_k \neq 0, \forall m \geq 1, k \leq m$  (序列  $x_i$  不是白噪声)  $\leftrightarrow$

$H_1: \rho_1 = \rho_2 = \dots = \rho_m = 0, \forall m \geq 1$  (序列  $x_i$  是白噪声)

$LB$  检验统计量为

$$Q_{LB} = n(n+2) \sum_{k=1}^m \left( \frac{\hat{\rho}_k^2}{n-k} \right)$$

使用 EVIEW6.0 可得数据的自相关系数和偏自相关系数图:

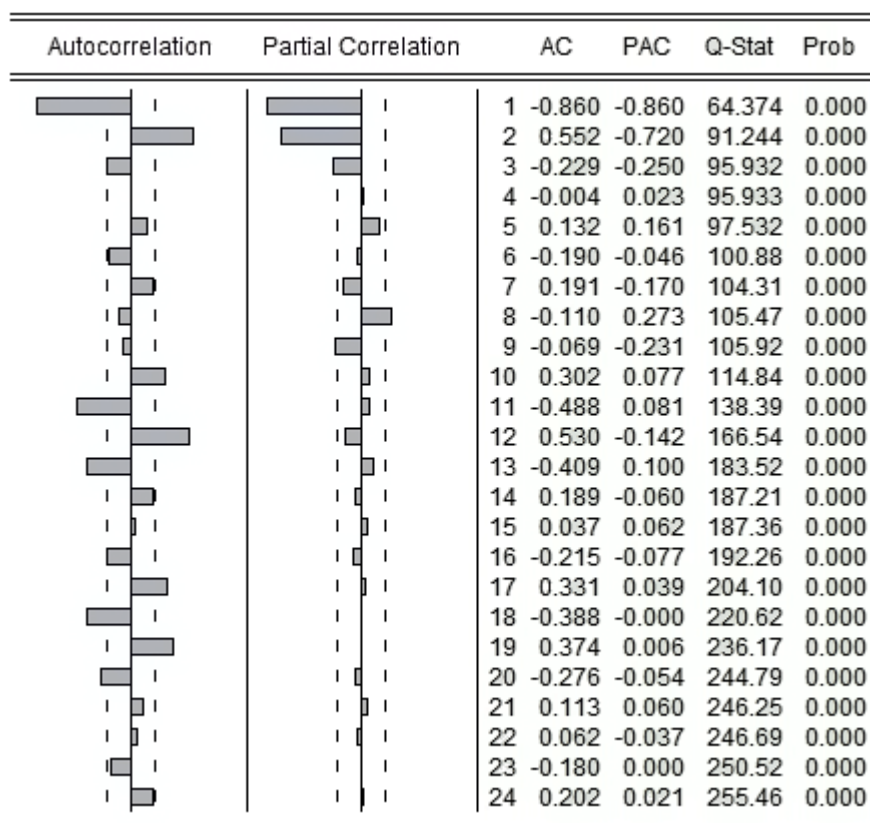


图 3.4 自相关系数和偏自相关系数图

由数据结果易知,  $Q$  统计量的值均不为零, 故接受原假设, 即认为数据不是白噪声序列, 满足建模条件。

## 3.3 模型构建

### 3.3.1 初选模型

根据前面的分析,  $SARIMA(p, d, q)(P, D, Q)^s$  中  $d = 0, D = 1, s = 12$ , 由于  $k=12$ , 时序列自相关系数和偏自相关系数都不显著为 0, 因此考虑  $P = 1, Q = 1$ 。观察序列自相关系数和偏自相关系数图发现, 自相关系数拖尾。而偏自相关系数表现为 2 阶截尾,  $k = 8, k = 9$  时有增大趋势, 因此考虑  $p = 0, q = 1$  或  $q = 2$ 。因此, 初选模型  $SARIMA(0, 0, 1)(1, 1, 1)^{12}$ ,  $SARIMA(0, 0, 2)(1, 1, 1)^{12}$ 。

### 3.3.2 参数估计

采用极大似然估计法对筛选出的模型进行参数估计, 结果如下:

模型	变量	参数	t值	P值	R-squared	调整R-squared	AIC	SC
SARIMA (0, 0, 1) (1, 1, 1) 12	AR (12)	1.280	50.867	0	0.2932	0.2727	12.1	12.2
	MA (1)	-0.524	-5.050	0				
	SMA (12)	-1.161	-13.191	0				
SARIMA (0, 0, 2) (1, 1, 1) 12	AR (12)	1.159	24.004	0	0.2219	0.1875	12.23	12.36
	MA (1)	-0.493	-4.019	0.0001				
	MA (2)	-0.048	-0.395	0.694				
	SMA (12)	-0.838	-18.607	0				

图 3.5 模型参数估计结果

$SARIMA(0, 0, 2)(1, 1, 1)^{12}$  没有通过显著性检验的模型, 并且 AIC 和 SC 值较大, 因此选择  $SARIMA(0, 0, 1)(1, 1, 1)^{12}$  作为最终模型。

### 3.3.3 残差分析

对  $SARIMA(0, 0, 1)(1, 1, 1)^{12}$  进行残差检验, 检验结果如下:

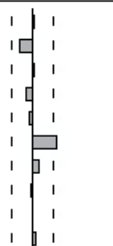

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1	0.019	0.019	0.0276
		2	-0.148	-0.149	1.7021
		3	0.008	0.015	1.7074
		4	-0.074	-0.099	2.1361
		5	-0.047	-0.041	2.3130
		6	0.268	0.252	8.1064
		7	0.072	0.051	8.5299
		8	-0.019	0.054	8.5588
		9	0.003	0.010	8.5595
		10	0.033	0.076	8.6550

图 3.6 残差检验

从图中可以看出, 除了  $k=6$ , 显著性为 4.4%, 小于 5%, 其他所有样本显著性检验均大于 5%, 并且落在 2 倍标准差内, 可以认为数据通过了残差检验。





## 4 基于神经网络的秦皇岛旅游接待人数的影响和预测回归模型

### 4.1 2020、2021 疫情对秦皇岛旅游业接待人数的影响程度

通过建立的  $SARIMA(0, 0, 1)(1, 1, 1)^{12}$  模型对 2020, 2021 每个月数据进行预测, 数据显示平均误差只有 8.8%, 模型效果较好。

时间	实际值	预测值	误差	时间	实际值	预测值	误差
2014-1	54.51	86.99	59.58%	2017-1	88.49	63.46	-28.29%
2014-2	31.77	27.01	-14.98%	2017-2	88.48	59.05	-33.26%
2014-3	45.80	37.64	-17.82%	2017-3	85.64	73.07	-14.68%
2014-4	112.10	64.22	-42.71%	2017-4	255.37	165.08	-35.36%
2014-5	227.68	250.06	9.83%	2017-5	465.29	430.02	-7.58%
2014-6	274.94	296.11	7.70%	2017-6	503.33	466.55	-7.31%
2014-7	643.20	718.44	11.70%	2017-7	924.53	969.23	4.84%
2014-8	678.13	733.73	8.20%	2017-8	1055.96	1055.18	-0.07%
2014-9	342.39	367.86	7.44%	2017-9	682.51	674.48	-1.18%
2014-10	297.26	311.75	4.87%	2017-10	775.54	552.74	-28.73%
2014-11	74.21	68.51	-7.67%	2017-11	189.86	47.74	-74.85%
2014-12	69.87	53.00	-24.14%	2017-12	139.15	32.63	-76.55%
2015-1	54.51	83.38	52.96%	2018-1	110.78	96.04	-13.30%
2015-2	62.21	37.78	-39.27%	2018-2	110.80	74.86	-32.44%
2015-3	60.63	43.66	-27.99%	2018-3	20.52	89.05	334.03%
2015-4	144.87	68.58	-52.66%	2018-4	20.52	256.69	1151.09%
2015-5	269.11	263.20	-2.20%	2018-5	972.19	738.82	-24.00%
2015-6	344.40	317.06	-7.94%	2018-6	625.63	483.20	-22.77%
2015-7	710.64	750.41	5.60%	2018-7	979.50	1078.24	10.08%
2015-8	758.71	748.91	-1.29%	2018-8	1290.62	1237.58	-4.11%
2015-9	437.89	384.56	-12.18%	2018-9	797.28	776.02	-2.67%
2015-10	348.90	313.70	-10.09%	2018-10	578.04	895.45	54.91%
2015-11	101.42	69.83	-31.15%	2018-11	578.04	393.66	-31.90%
2015-12	28.40	58.97	107.65%	2018-12	167.32	59.29	-64.56%
2016-1	69.36	114.29	64.77%	2019-1	128.42	130.40	1.54%
2016-2	78.30	78.79	0.64%	2019-2	128.44	107.64	-16.20%
2016-3	70.85	75.03	5.90%	2019-3	23.79	27.66	16.29%
2016-4	206.46	110.75	-46.36%	2019-4	23.79	-45.60	-291.72%
2016-5	359.12	311.54	-13.25%	2019-5	1127.03	1170.33	3.84%
2016-6	407.00	380.26	-6.57%	2019-6	671.43	781.39	16.38%
2016-7	804.20	845.77	5.17%	2019-7	1116.81	1308.66	17.18%
2016-8	932.55	848.08	-9.06%	2019-8	1514.50	1569.90	3.66%
2016-9	580.86	459.96	-20.81%	2019-9	1001.23	980.78	-2.04%
2016-10	474.96	343.27	-27.73%	2019-10	1003.23	695.90	-30.63%
2016-11	137.40	52.03	-62.14%	2019-11	288.76	507.09	75.61%
2016-12	97.00	-14.71	-115.16%	2019-12	234.90	304.37	29.57%
平均误差:				8.80%			

图 4.1 预测数据与真实数据误差

通过  $SARIMA(0, 0, 1)(1, 1, 1)^{12}$  模型预测秦皇岛 2020 年与 2021 年接待游客总数, 同时与实际接待游客总数, 可以得到疫情对接待总人数的影响。其中 影响百分比 =  $\frac{\text{影响值}}{\text{预测数据}}$ , 代表了疫情时接待游客总人数减少的比例。结果如下:



时间	实际数据	预测数据	影响值	影响百分比
2020-1	29.66	255.34	225.68	88.38%
2020-2	29.65	125.68	96.03	76.41%
2020-3	29.66	45.11	15.45	34.25%
2020-4	104.53	-54.94	-159.47	290.26%
2020-5	0.50	1417.69	1417.19	99.96%
2020-6	69.91	831.40	761.49	91.59%
2020-7	405.79	1448.49	1042.70	71.99%
2020-8	589.60	1748.81	1159.21	66.29%
2020-9	336.78	1204.91	868.13	72.05%
2020-10	560.30	1203.21	642.91	53.43%
2020-11	218.89	358.69	139.80	38.98%
2020-12	304.84	269.35	-35.49	-13.17%
2021-1	63.47	375.59	312.12	83.10%
2021-2	63.45	122.14	58.69	48.05%
2021-3	63.48	72.42	8.94	12.35%
2021-4	277.88	-155.72	-433.60	278.45%
2021-5	527.10	1789.80	1262.70	70.55%
2021-6	355.88	1036.19	680.31	65.65%
2021-7	703.40	1873.10	1169.70	62.45%
2021-8	164.95	2048.77	1883.82	91.95%
2021-9	164.95	1465.67	1300.73	88.75%
2021-10	223.73	1459.24	1235.51	84.67%
2021-11	262.10	448.22	186.12	41.52%
2021-12	252.55	313.46	60.91	19.43%
平均影响:				79.89%

图 4.2 模型预测结果及其影响

数据显示,受疫情影响,2020 年秦皇岛旅游接待总人数减少 80.87%,2021 年秦皇岛旅游接待总人数减少 78.91%,两年平均减少 79.89%。

从数据来看,2002 年新冠开始之初影响最大,5 月份近乎夭折。但随着武汉 4 月全面解封,2020 年疫情防控整体良好的情况下,2020 年下半年疫情的影响程度逐渐减小。2021 年由于德尔塔变异毒株传入中国,各地疫情零散爆发,2021 年影响整体平稳偏高,到年底影响逐渐减小。

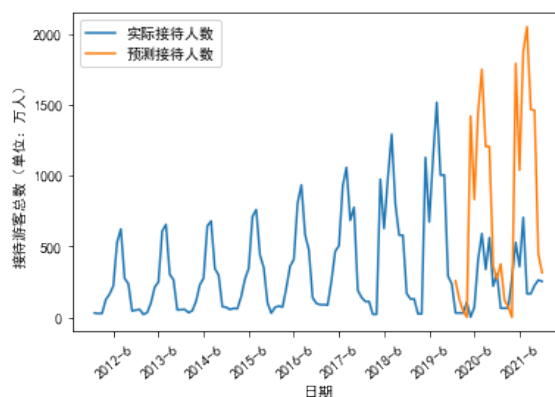


图 4.3 模型预测结果及其影响

## 4.2 预测 2022 年疫情对秦皇岛旅游业接待人数影响程度

### 4.2.1 基于 lstm 神经网络构建回归预测模型

为了能够对 2022 年疫情影响旅游业情况进行预测，选取 2012 年与 2019 年真实的数据作为数据集训练 LSTM 神经网络，其中 2012 年-2018 年为训练集，2019 年为测试集。数据集具体划分如下：

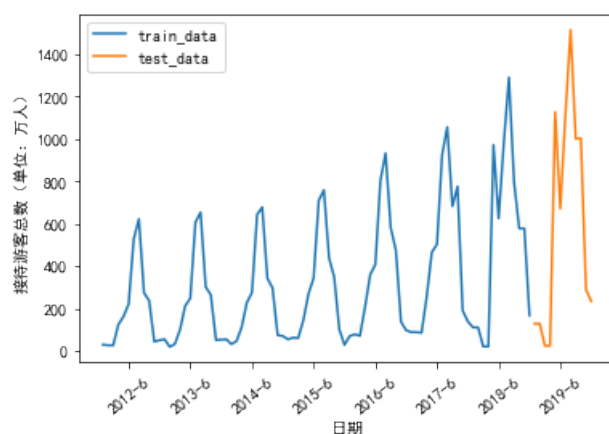


图 4.4 数据集划分

其中参数  $n$  为 24，即认定当前一个月的数据主要受过去 24 个月数据影响。搭建 LSTM 神经网络进行训练，训练次数为 2500 次，均方差随训练次数的变化如下图所示：

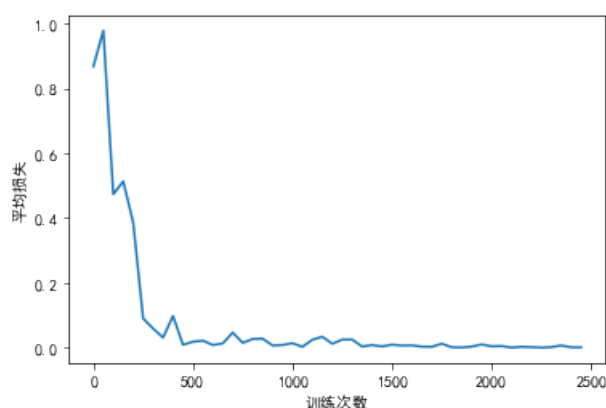


图 4.5 损失值变化

损失值随着训练次数增多逐渐下降并波动，最终趋向于 0，说明模型在训练集上拟合效果良好。使用模型在测试集上进行训练，模型回归效果如下：

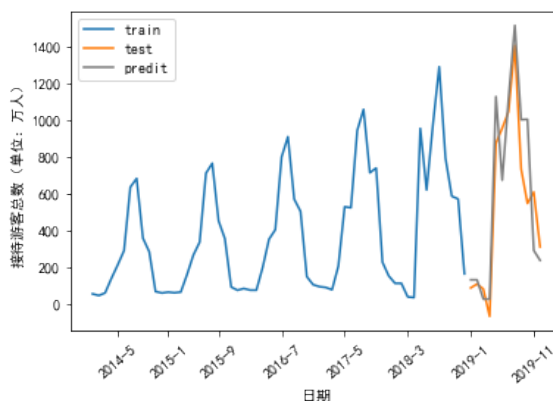


图 4.6 模型在测试集上的表现

可以看出，在测试集上，实际曲线与预测曲线基本重合，模型在测试集上回归效果良好。

## 4.2.2 基于回归模型对 2022 年影响程度进行预测

基于上面建立的 LSTM 模型，使用 2020 年，2021 年的数据，即疫情后 24 个月的数据，逐步预测 2022 年 12 个月秦皇岛接待游客的总人数，预测结果图示如下：

时间	ARIMA预测	LSTM预测	影响值	影响百分比	较2021增长
2022-1	529.54	142.37	387.18	73.12%	124.30%
2022-2	117.61	124.49	-6.88	-5.85%	96.20%
2022-3	107.38	46.44	60.94	56.76%	-26.85%
2022-4	400.00	157.16	242.84	60.71%	-43.44%
2022-5	2266.19	273.43	1992.75	87.93%	-48.13%
2022-6	1298.36	410.41	887.95	68.39%	15.32%
2022-7	2416.70	746.69	1670.00	69.10%	6.15%
2022-8	2432.79	478.97	1953.82	80.31%	190.38%
2022-9	1799.50	314.30	1485.19	82.53%	90.55%
2022-10	1787.01	359.98	1427.03	79.86%	60.90%
2022-11	562.84	142.80	420.04	74.63%	-45.52%
2022-12	369.93	154.74	215.18	58.17%	-38.73%
平均:	1173.99	279.32	894.67	65.47%	31.76%

图 4.7 2022 年预测结果

从数据可以看出，疫情对秦皇岛 2022 年旅游业的冲击依然很大，相比疫情前，2022 年秦皇岛接待游客总数相比疫情前减少 65.47%，但是比 2021 年减少 14.42%，说明疫情对秦皇岛的旅游影响变小，秦皇岛旅游业正逐步变暖。2022 年秦皇岛接待游客总量预计比 2021 年增长 31.76%，符合刚才的判断。

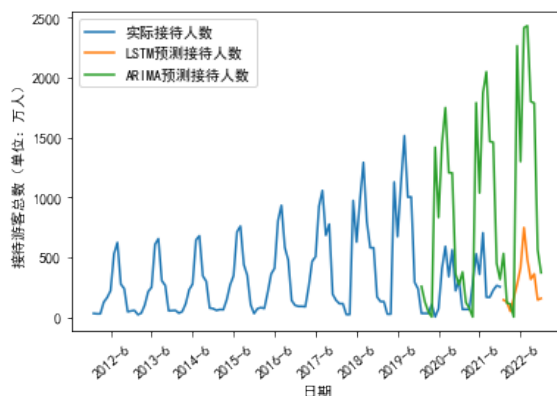


图 4.8 2022 年预测结果对比

绘制曲线图，任然可以得出相似的结论，从曲线图中可以发现，疫情对秦皇岛 2022 年旅游业的冲击相对疫情前依然很大，但相比 2020 与 2021 年，曲线呈上升趋势，说明疫情对秦皇岛的旅游影响变小，秦皇岛旅游业正逐步变暖。

## 5 主要结论和存在问题

### 5.1 主要结论

根据以上分析，主要得出以下结论：

**(1)** 受疫情影响，2020 年秦皇岛旅游接待总人数减少 80.87%,2021 年秦皇岛旅游接待总人数减少 78.91%, 两年平均减少 79.89%。**(2)** 预计疫情对秦皇岛 2022 年旅游业的冲击依然很大，相比疫情前减少 65.47%。**(3)** 相比 2020 与 2021 年，疫情对秦皇岛的旅游影响变小，秦皇岛旅游业相比 2020，2021 呈上升趋势，预计比 2021 年增长 31.76%。

### 5.2 存在问题

**(1) 数据缺失** 搜集数据时发现缺失了一部分数据，由于数据的缺失，导致模型的预测精确度下降，预测结果出现了少量负值。由于缺失 2022 年的旅游数据，因此无法对未来时段秦皇岛的旅游发展做出更准确的预测。

**(2)Omicron 的影响** 2022 年新冠病毒变异株 Omicron 输入中国境内，2 月秦皇岛毗邻城市葫芦岛疫情，3 月长春疫情，4 月上海疫情再次对秦皇岛造成了冲击。由于缺失 2022 年的秦皇岛旅游数据，无法将 Omicron 的影响考虑到模型中。



## 参考文献

- [1] 邓集贤, 杨维权, 司徒荣, 邓永录. 概率论与数理统计. 下册 (第四版). 北京: 高等教育出版社, 2009.7
- [2] 财务视角下新冠肺炎疫情对贵州工业发展影响的统计测度 [C]//中国统计教育学会.2020 年 (第七届) 全国大学生统计建模大赛优秀论文集.[出版者不详],2020:25.DOI:10.26914/c.cnkihy.2020.045581.
- [3] 新冠肺炎疫情对青海省旅游业影响的统计研究——基于 SARIMA 模型 [C]//中国统计教育学会.2020 年 (第七届) 全国大学生统计建模大赛优秀论文集.[出版者不详],2020:16.DOI:10.26914/c.cnkihy.2020.045584.
- [4] 周志华. 机器学习 [J]. 清华大学出版社, 2016, 8(28): 1-415.
- [5] [1]. 新冠疫情对海南旅游业影响的统计测度研究 [C]//中国统计教育学会.2020 年 (第七届) 全国大学生统计建模大赛优秀论文集.[出版者不详],2020:29.DOI:10.26914/c.cnkihy.2020.045597.
- [6] 王燕. 应用时间序列分析. 中国大学出版社, 2005.7