|  |  |  |
| --- | --- | --- |
| **Problem Chosen** C | **2024 MCM/ICM Summary Sheet** | **Team Control Number** 2410809 |

summary

【英文字体大小】

【右上角的页码从哪里开始编号】

【正文里答案可以标粗】

【可以用“我们”】

【总分总写法】

【对图的分析要充分】

【摘要第一段写一下背景，3-4行】

每段开头：Firstly Secondly Next Besides Finally

**Keybord:**

**Content**

[summary 1](#_Toc158021123)

[1 Introduction 3](#_Toc158021124)

[1.1 Background of the Problem 3](#_Toc158021125)

[1.2 Restatement of the Problem 3](#_Toc158021126)

[1.3 Literature review（文献综述） 4](#_Toc158021127)

[1.4 Our Work（问题分析的精简版） 4](#_Toc158021128)

[2 Assumptions and Notations（假设和符号） 6](#_Toc158021129)

[2.1 Assumptions（问题假设） 6](#_Toc158021130)

[2.2 Natations(符号说明) 6](#_Toc158021131)

[3 Data processing 7](#_Toc158021132)

[3.1 Processing missing value 7](#_Toc158021133)

[3.2 Handling of outliersm 8](#_Toc158021134)

[4 Task 1 Capture game flow 9](#_Toc158021135)

[4.1 建立势头模型 10](#_Toc158021136)

[4.2 确定 10](#_Toc158021137)

[4.3 选择平滑系数 13](#_Toc158021138)

[4.4 可视化比赛流程 14](#_Toc158021139)

[5 Task 2 探究势头的作用 15](#_Toc158021140)

[5.1 探究相减势头的时间序列是否随机 15](#_Toc158021141)

[5.2 讨论势头与比赛结果的关系 17](#_Toc158021142)

[5.3 小结 19](#_Toc158021143)

[6 Task 3 探究比赛流程的转变 19](#_Toc158021144)

[6.1 预测波动 20](#_Toc158021145)

[6.2 寻找与波动最相关的因素 22](#_Toc158021146)

[6.3 提供建议 24](#_Toc158021147)

[7 Task 4+解决的问题 25](#_Toc158021148)

[7.1 aaaaaaaaaaaa 25](#_Toc158021149)

[7.2 aaaaaaaaaaaa 25](#_Toc158021150)

[8 备忘录 25](#_Toc158021151)

[9 Sensitivity Analysis（敏感性分析） 25](#_Toc158021152)

[10 Model Evaluation and Further Discussion 25](#_Toc158021153)

[10.1 Strength（优点） 25](#_Toc158021154)

[10.2 Weaknesses（缺点） 25](#_Toc158021155)

[10.3 Further Discussion（推广） 26](#_Toc158021156)

[References【单独一页】 27](#_Toc158021157)

# 1 Introduction

## 1.1 Background of the Problem

In the 2023 Wimbledon World Cup final, 20-year-old Spanish rising star Carlos Alcascaras defeated 36-year-old Novak Djokovic. Djokovic, one of the greatest players in Grand Slam history, lost for the first time.

Here's a recap of how they played. Djokovic won the first set 6-1, Alcalas won the second set 7-6, and the next set 6-1. But it was Djokovic who won the fourth set 6-3. So far, Alkaras has settled for a draw with Djokovic. In the early stages of the fifth set, Djokovic held a commanding 4-0 lead, but from the fifth game, Alkaras stormed back to take the fifth set 6-4. So the final winner of this game is Alcaras.

As can be seen from the above, there have been many "momentum" shifts in this game. The dictionary definition of momentum is " strength or force gained by motion or by a series of events. "; In sports, it means "strength." Often manifested as a phenomenon that is difficult to measure, a team or player may feel that they have momentum during a game . However, the influence of various events in the game on the "momentum" shift is not obvious.

## 1.2 Restatement of the Problem

Given the context as well as every point data after the first two rounds of the men's competition at Wimbledon 2023, the following questions need to be addressed in this paper:

* Task 1 required building a model that would allow it to identify the better performers at a particular time in the race, as well as reflect how well they performed. And show the flow of the game in a visual way.*Note: Players tend to perform better when serving, so this can be factored into the model.*
* Task 2 asks to evaluate a coach's view, based on a model, that game fluctuations are random and that "momentum" does not play a role in a game.
* Task 3 requires using data from at least one match to develop a model that enables it to predict turning points in a match, that is, which events will affect the direction of the outcome of the match, and analyze which factors are associated with it. And take into account the differences in the "momentum" fluctuations of the game, to provide suggestions for a new game.
* Task 4 requires testing the model to assess the accuracy of the model's predictions and its generalization to other sports or competitions (such as Women’s matches , tournaments, table tennis, etc.).
* Task 5 asked to write a memo, includes a summary of the results, suggestions for coach how to use "momentum", about players affect the process of incident response.

## 1.3 Literature review（文献综述）

## 1.4 Our Work（问题分析的精简版）

【放思维导图+一些解释（背景色要淡，字不能太小）】

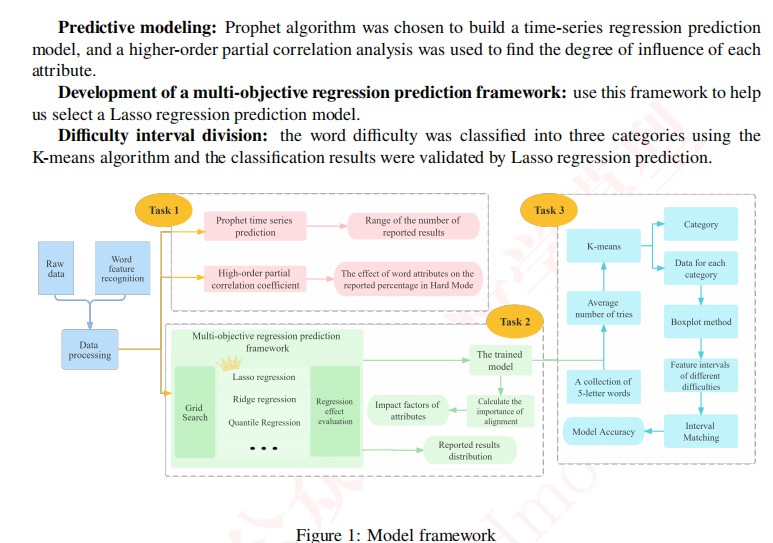
* 2317本文接下来的内容安排如下。在第二节中，我们介绍了前提假设和理由，并在第三节中提到了公式中的常见变量。第四部分是建模前的数据预处理。第五节建立报告编号区间预测模型，并探讨词属性与模式选择之间的关系。第六节，建立报告结果分布预测模型。第七节，我们提出了单词难度分类模型。第八节继续探讨数据文件的有趣特征。第九节和第十节分析了模型的灵敏度，并进一步评估了模型的优缺点。最后，第十一节给出结论。
* 2351基于对问题的分析，我们提出了图 1 所示的模型框架，主要由以下几部分组成：

**数据分析**：处理报告数据并确定词语的特征。

**预测模型**：选择 Prophet 算法建立时间序列回归预测模型，并使用高阶偏相关分析找出各属性的影响程度。

**开发多目标回归预测框架**：利用该框架帮助我们选择 Lasso 回归预测模型。

**难度区间划分**：使用 K-means 算法将单词难度分为三类，并通过 Lasso 回归预测对分类结果进行验证。



* 2336【可以学习，很模板化】

首先，我们构建了可以衡量词语熟悉程度、构成特征、关联程度和混淆程度的四类指标，并用这些指标来反映词语的属性。【针对问题 1，我们基于 SIR 模型建立了一个名为 "目标-两名玩家-流失（T2PL）"的动态系统，以解释 Wordle 报告的每日波动。此外，玩家还被分为两类：普通玩家和忠诚玩家，每类玩家的流失率各不相同。这样，模型就能更好地模拟不同时间段内不平等的流失率。我们还探讨了单词属性与 "困难模式 "玩家数量之间的关系，发现某些属性会影响 "困难模式 "报告的百分比。】

其次，我们在 SIR 模型的基础上开发了 T2PL 模型，该模型是一个动态模型，可以很好地解释所报告结果的数量和所报告的 "困难模式 "结果所占百分比的总体趋势。在此基础上，我们探讨了单词属性对所报告的 "困难模式 "结果百分比的影响。【针对问题 2，我们开发了一个 P&S 模型，该模型使用模拟算法和梯度下降法来模仿玩家猜词和分享游戏结果的行为。模拟器的工作原理是利用可观察到的信息剔除所有不满意的单词，然后以单词频率为权重从剩余的单词列表中随机抽取单词。然而，我们发现模拟结果与真实分布并不完全一致。因此，我们用 7 个变量重新调整了分布，这 7 个变量代表了玩家在获得不同分数时分享分数的可能性。通过梯度下降法对它们进行优化，可以生成更好的分布预测。使用 P&S 模型，我们预测 EERIE 一词在 2023 年 3 月 1 日的分布为（0, 0, 9%, 29%, 45%, 14%, 3）】

第三，我们利用该算法模拟了wordle玩家猜词时的策略，从而模拟了结果的初始分布。考虑到玩家的心理特点，我们添加了表示玩家分享分数意愿的参数，并模拟了报告结果的最终分布。【对于问题 3，我们需要按难度对谜题进行分类。我们使用 3 个聚类 K-means 对所有报告的试题分布进行聚类分析，每个聚类分为简单、中等和困难。我们使用随机森林模型，利用开头定义的属性指标将谜题分为这三个类别。我们计算了每个指标与难度之间的相关系数，以显示这些指标对谜题难度的影响方向。同时还讨论了聚类的敏感性。根据我们的模型，EERIE 的难度为难】

第四，我们根据得分分布对单词进行聚类，并根据难度将单词分为三类。聚类结果作为标签，用于构建随机森林模型，根据词的属性对词的难度进行分类。【对于问题 4，我们进一步探讨了单词难度的影响。通过线性回归，我们发现单词难度对报告结果的数量有明显的影响：较难的谜题会导致较少的报告。如前所述，难度还与选择 "困难模式 "的人数比例相关。通过这部分研究，我们发现相关性是由单词难度影响普通模式玩家人数而形成的。】

最后，根据上述模型的结果，我们进行了进一步的探索，发现了一些有趣的结论。【通过发现单词属性、谜题难度和游戏报告模式之间的相互作用，Wordle 操作员可以更深入地了解玩家。基于这一发现，我们还可以提出一些合理的建议。】

# 2 Assumptions and Notations（假设和符号）

## 2.1 Assumptions（问题假设）

·**Assumption 1.**-----

Reason 1:----

·**Assumption 2.**-----

Reason 2:---

·**Assumption 3.**-----

Reason 3:---

我们好像假设了累计相减势头作为看比赛波动的指标应该

然后假设什么场地什么其他我们没发处理的因素对比赛走势不影响

## 2.2 Natations(符号说明)

Table Nations

|  |  |
| --- | --- |
| Symbol | Definition |
|  | 玩家*k*在时刻下的势头 |
|  | 开始换局或休息的时间点 |
|  | serve\_no |
|  | server |
|  | winner\_shot\_type |
|  | pk\_double\_fault |
|  | pk\_unf\_err |
|  | pk\_net\_pt |
|  | pk\_break\_pt |
|  | p1\_distance\_run与rally\_count的比值 |
|  | speed\_mph与100的比值 |
|  | (p1\_points\_won-p2\_points\_won)与(p1\_points\_won+p2\_points\_won)的比值 |
|  | pk\_ace |
|  |  |
|  | p1\_points\_won |
|  | P2\_points\_won |

# 3 Data processing

In order to make the data better adapt to the requirements of the model and the characteristics of the algorithm, we usually need to preprocess the data before formally solving the problem.

## 3.1 Processing missing value

There may be missing values in the raw data, and a heat map can help us detect this, visualizing the distribution of missing values.A heatmap of the raw data is presented below.

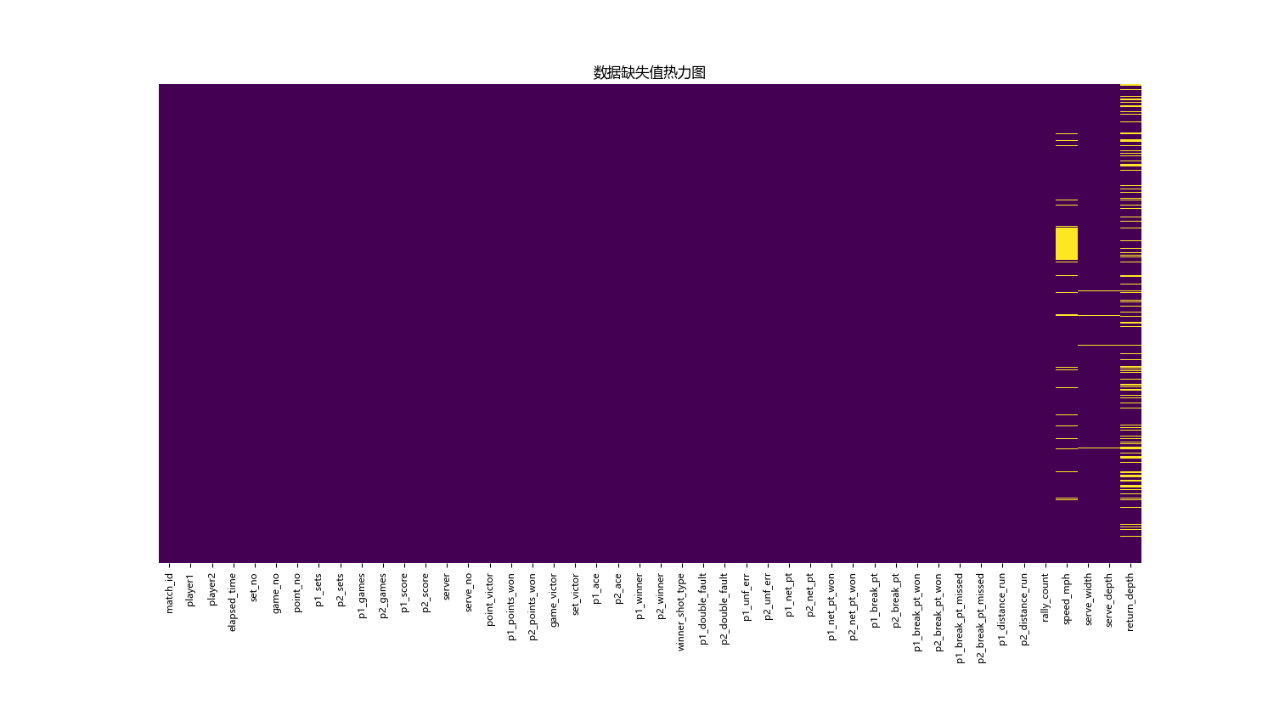


图 Heatmap of missing data

In the graph, purple represents non-missing values and yellow represents missing values. As can be seen from the figure, the data of most indicators are complete, but some indicators obviously have missing values. Further, we can see that a whole block of "speed\_mph" is missing, "serve\_width, serve\_depth" has fewer missing values, and "return\_dept" has more dense missing values.

In order to improve the quality of the data and ensure that the model can make better use of the information, we fill in the missing values, as follows:

* **For speed\_mph**: After looking at the table data based on the above results, we found a whole block of missing values in lines 2187 to 2674, namely Daniel Elahi Galan vs. Mikael Ymer and Guido Pella vs. Roman Safiullin. Therefore, the average of players with similar rankings can be used instead. For the rest, we will fill forward or backward using numerical time series.
* **For serve\_width**: serve\_depth: Since the missing values of these two indicators are less, the mode "C" and "NCTL" are used to fill.
* **For return\_dept**: The original data only has D and ND, so for missing values, we can introduce "MD" (medium depth) instead.

According to the above method, the data obtained after processing is shown in the following table (only part is shown here) :

表 Data processed with missing values (part)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| match\_id | player1 | player2 | **speed\_mph** | **serve\_width** | **serve\_depth** | **return\_dept** |
| 2023-wimbledon-1301 | Carlos Alcaraz | Nicolas Jarry | 125 | C | NCTL | MD |
| 2023-wimbledon-1302 | Alexander Zverev | Matteo Berrettini | 106 | C | NCTL | MD |
| … | | | | | | |
| 2023-wimbledon-1310 | Daniel Elahi Galan | Mikael Ymer | 115 | C | NCTL | MD |
| 2023-wimbledon-1311 | Guido Pella | Roman Safiullin | 115 | C | NCTL | MD |
| … | | | | | | |

## 3.2 Handling of outliersm

(1) p1\_score and p2\_score

There may also be outliers in the raw data. On a preliminary analysis of the raw data, we found that the p1\_score and p2\_score fields should contain only a few specific values (0,15, 30, 40, AD), and the rest should be treated as outliers.

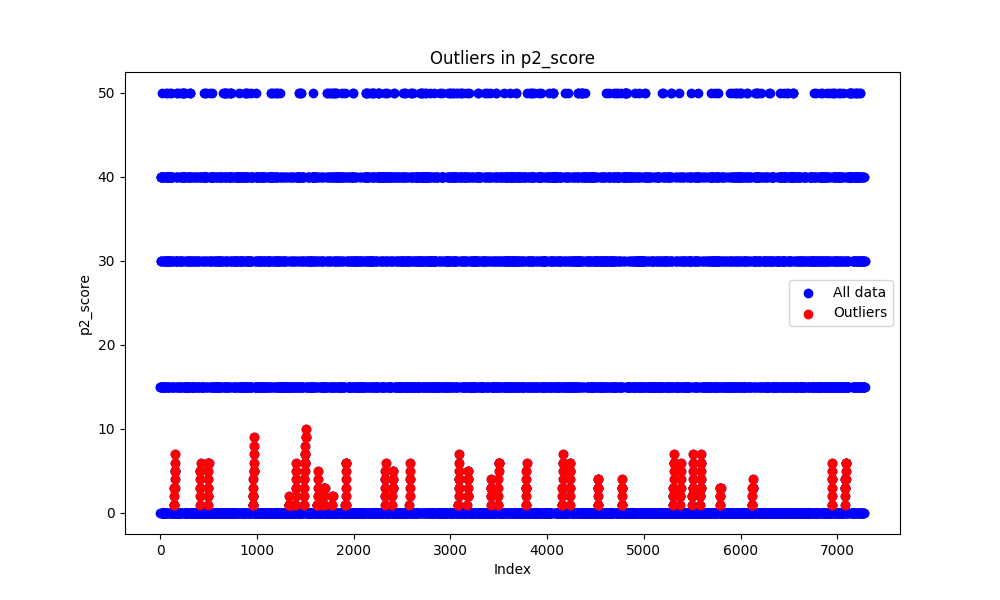
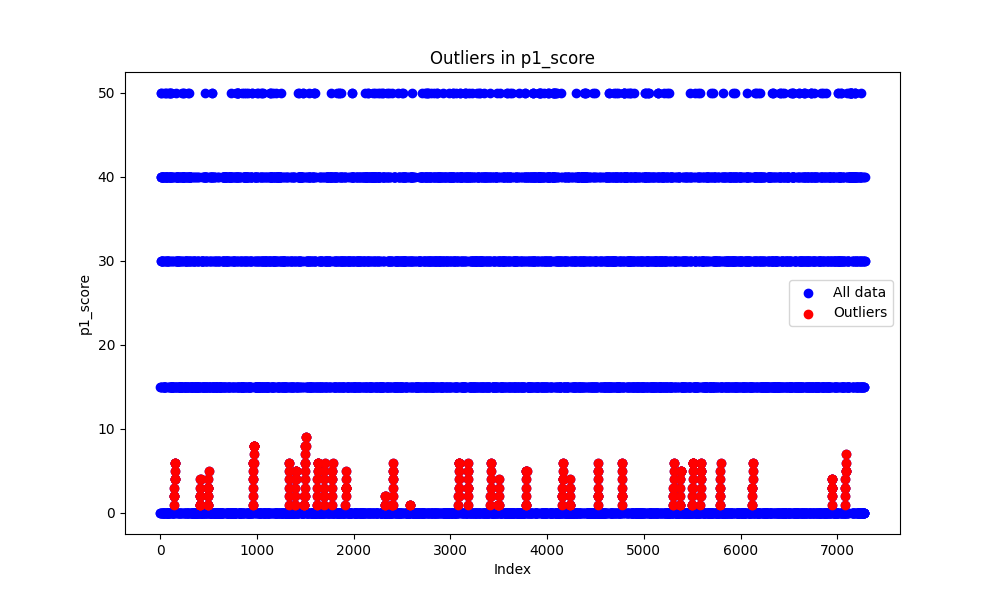
 Therefore, we can use rule-based anomaly detection to analyze the data, which is especially suitable for cases where there is a clear set of valid values in the data. At the same time, we define AD as 50, that is, we define specific values as (0,15, 30, 40, 50). Below is the test result.

图 Outlier detection results for p1\_score and p2\_score

As can be seen from the figure, in all the data, the p1\_score and p2\_score fields have outliers between 0 and 10.

However, considering that the data of these two features have little significance for the following modeling, we choose to delete all the data of these two features.

(2) p1\_points\_won和p2\_points\_won

In fact, the outliers in this part were discovered by accident in the subsequent problem solving process. In order to ensure the consistency of the structure of the paper and avoid the influence of the bad outliers, we choose to give corresponding treatment here.

表 Outliers of the original data

|  |  |  |
| --- | --- | --- |
| Line | p1\_points\_won | p2\_points\_won |
| 4055 | 12 | 8 |
| 4056 | 21 | 21 |

It can be seen from the table that the values of 12 and 21 and 8 and 21 in these two rows are not continuous, but mutated. Based on the two metrics "p1\_points\_won" and "p2\_points\_won" which mean "the cumulative score of the player", we know that there is an outlier.

Combining the actual meaning of the two indicators, as well as the actual meaning and data of other indicators in the original data, we can easily deduce the correct value. The following shows the results after processing.

表 Processed p1\_points\_won and p2\_points\_won

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| line | p1\_points\_won | p2\_points\_won |  | line | p1\_points\_won | p2\_points\_won |  | line | p1\_points\_won | p2\_points\_won |
| 4035 | 12 | 9 |  | 4043 | 16 | 13 |  | 4051 | 18 | 19 |
| 4036 | 12 | 10 |  | 4044 | 16 | 14 |  | 4052 | 18 | 20 |
| 4037 | 12 | 11 |  | 4045 | 17 | 14 |  | 4053 | 19 | 20 |
| 4038 | 13 | 11 |  | 4046 | 17 | 15 |  | 4054 | 20 | 20 |
| 4039 | 14 | 12 |  | 4047 | 17 | 16 |  | 4055 | 21 | 20 |
| 4040 | 14 | 12 |  | 4048 | 17 | 17 |  | 4056 | 21 | 21 |
| 4041 | 14 | 13 |  | 4049 | 18 | 17 |  |  |  |  |
| 4042 | 15 | 13 |  | 4050 | 18 | 18 |  |  |  |  |

As can be seen from the table, outliers appear in lines 4035-4055. Further, we can see that during this period, player 1 was first ahead of Player 2 by 2 or 3 points, but was gradually evened by Player 2.

# 4 Task 1 Capture game flow

在这一部分，我们首先建立了势头的EWMA模型，用于捕捉得分发生时的比赛流程。然后根据模型分别求出球员1和球员2随比赛流程变化的势头，并根据两者的相减势头来比较它们的表现。

## 4.1 建立势头模型

EWMA模型是一种时间序列分析中常用的平滑方法，用于捕捉数据的趋势和变化。

在本题中，为了更好地理解玩家在比赛中的表现变化，我们将势头视为因变量，并将某一时刻的势头定义为“该时刻下的技术表现”加上“前一时刻的势头”。具体而言，“该时刻下的技术表现”与“双误、开球制胜、非受迫性失误、以及击发不可触及的一球”等指标相关。

根据上述分析，我们建立的指数加权移动平均模型是这样的：时刻下玩家*k*的势头为

其中，是平滑系数，通常取值在0到1之间，表示对的权重，值越大表示越重视。是“时刻下的技术表现”，也是根据比赛事件计算的原始得分。是玩家*k*在时刻下的势头。

因为与多个技术指标相关，因此可表示为

其中，为技术指标。

根据上述的分析，我们容易得到，为了确定时刻下的势头，关键就是确定和。下面展开详细说明。

## 4.2 确定

4.2.1 技术指标的选取

【微调一下】

在网球比赛中，不同时刻呈现不同的技术表现，这些技术表现在影响比赛势头方面发挥着重要作用。在所提供的比赛数据中，共列出了十多项单打比赛的技术统计指标。为了创建一个平衡的特征集，涵盖技术表现的不同方面，我们给出了下面这12个特征。

为首次发球是否失误。失误记为1，无失误记为0。这个特征捕捉了球员开始比赛时的能力。

为是否作为发球方，因为发球方通常具有战略优势。

为不可触及的镜头类别，这提供了关于球员技能和打法风格的见解。

为双误。双误是重大失误，可能导致失分，监控这一点有助于评估球员的一致性。

为非受迫性失误，表明球员的控制和决策能力。

为成功上网次数，用于展示球员的积极性和结束比赛的能力。

为破发点，说明球员有机会打破对手发球局，这是网球比赛中至关重要的一环。

为跑动距离与射击次数比值，体现了球员在比赛中的移动效率。

为发球速度。

为得分差异比值，它捕捉了球员在赢得比赛中相对于对手的优势，提供了相对度量。

为Ace，展示了球员轻松获得分数的能力。

为短期得分率，是球员当前状态和即时比赛表现的有价值的度量。以下给出具体的数学表达：

规定这个短期是5个回合（考虑了休息时间和换场时间），记为

其中，表示“p1\_points\_won”，表示“p2\_points\_won”，表示开始换局或休息的时间点。

对于12个指标进行加权求和，便可得到时刻下根据比赛事件计算的原始得分

其中，分别表示的权重。但是这些权重暂时也是未知的，因此下面我们来解决这个问题。

4.2.2 权重的确定

层次分析法（Analytic Hierarchy Process，AHP）是一种多标准决策分析方法，用于确定不同因素或标准的相对重要性。

对于技术表现来说，选取的12个特征具有不同的重要性，因此选择层次分析法来量化这些特征的重要性。下面是运用该方法的具体步骤。

***Step1***:建立系统的递阶层次结构

将决策问题分解为三个层次，最上层为目标层M；中间层为准则层，包括重要性C1、常见性C2、实时性C3、稳定性C4、相关性C5 五个因素；最下层为方案层，即12个特征。如下图所示。

【层次结构图】

***Step2***:构造判断矩阵

将决策层C中的特征两两比较，得到判断矩阵如下所示，

表 准则层5个因素的判断矩阵

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 重要性C1 | 常见性C2 | 实时性C3 | 稳定性C4 | 相关性C5 |
| 重要性C1 | 1 | 7 | 5 | 7 | 5 |
| 常见性C2 | 1/7 | 1 | 1/2 | 1 | 1/2 |
| 实时性C3 | 1/5 | 2 | 1 | 2 | 1 |
| 稳定性C4 | 1/7 | 1 | 1/2 | 1 | 1/2 |
| 相关性C5 | 1/5 | 2 | 1 | 2 | 1 |

对于每个因素，将12个特征两两比较，得到5个判断矩阵。下面展示对于重要性12个特征的判断矩阵。

表 判断矩阵

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C1 |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 1 | 1/3 | 1/9 | 1/7 | 3 | 1/3 | 1/3 | 2 | 1 | 2 | 1/3 | 2 |
|  | 3 | 1 | 1/3 | 1/7 | 6 | 1/2 | 1/2 | 3 | 2 | 5 | 1/2 | 5 |
|  | 9 | 3 | 1 | 3 | 7 | 5 | 1 | 7 | 3 | 9 | 3 | 9 |
|  | 7 | 7 | 1/3 | 1 | 9 | 1/3 | 1 | 7 | 2 | 9 | 3 | 7 |
|  | 1/3 | 1/6 | 1/7 | 1/9 | 1 | 1/5 | 1/7 | 1 | 1/9 | 3 | 1/9 | 2 |
|  | 3 | 2 | 1/5 | 3 | 5 | 1 | 1/7 | 1 | 2 | 7 | 1/5 | 5 |
|  | 3 | 2 | 1 | 1 | 7 | 7 | 1 | 7 | 2 | 9 | 1 | 7 |
|  | 1/2 | 1/5 | 1/9 | 1/7 | 3 | 1/5 | 1/7 | 1 | 1/3 | 3 | 1/7 | 2 |
|  | 1 | 1/2 | 1/3 | 1/2 | 9 | 1/2 | 1/2 | 3 | 1 | 9 | 1/2 | 2 |
|  | 1/2 | 1/5 | 1/9 | 1/9 | 1/3 | 1/7 | 1/9 | 1/3 | 1/9 | 1 | 1/9 | 2 |
|  | 3 | 2 | 1/3 | 1/3 | 9 | 5 | 1 | 7 | 2 | 9 | 1 | 6 |
|  | 1/2 | 1/5 | 1/9 | 1/7 | 1/2 | 1/5 | 1/7 | 1/2 | 1/6 | 1/2 | 1/6 | 1 |

***Step3***:一致性检验

在使用判断矩阵求权重之前，必须对上述的6个判断矩阵都进行一致性检验。通常，我们对一致性指标以及一致性比例进行计算，若一致性比例小于0.10，则认为该判断矩阵一致性可以接受。其中，一致性指标，一致性比例。

经过计算，我们容易得到~这六个判断矩阵的一致性检验结果，

表 ~的一致性检验结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | 0.0051 | 0.1420 | 0.0579 | 0.1146 | 0.0668 | 0.0509 |
|  | 0.0046 | 0.0922 | 0.0376 | 0.0744 | 0.0434 | 0.0330 |

我们可以看到，~的一致性比例均小于0.20，因此它们的一致性均可以接受.

***Step4***:权重的求解

接下来使用判断矩阵求权重。为了保证结果的稳健性，我们分别采用了算术平均法、几何平均法、特征值法这三种方法求出了权重，再计算上述结果的平均值，作为最后的权重。下面展示使用判断矩阵求权重的结果：

表 根据求权重的结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 算数平均法 | 几何平均法 | 特征值法 | 平均值 |
| 重要性C1 | 0.584 | 0.5851 | 0.5864 | 0.5852 |
| 常见性C2 | 0.0727 | 0.0725 | 0.0723 | 0.0725 |
| 实时性C3 | 0.1353 | 0.1350 | 0.1345 | 0.1349 |
| 稳定性C4 | 0.0727 | 0.0725 | 0.0723 | 0.0725 |
| 相关性C5 | 0.1353 | 0.1350 | 0.1345 | 0.1349 |

根据上表，我们认为重要性、常见性、实时性、稳定性、相关性这五个因素在之间的权重分别是0.585167，0.0725，0.134933，0.0725，0.134933。

同上面的方法，我们可以得到对于每一个因素12个特征之间的权重，整理的数据如下表所示，

表 最终的权重表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 指标权重 |  |  |  |  |  |  |
| C1 | 0.5852 | 0.0352 | 0.0705 | 0.2178 | 0.1600 | 0.0192 | 0.0884 |
| C2 | 0.0725 | 0.0943 | 0.1481 | 0.0740 | 0.1588 | 0.0666 | 0.0666 |
| C3 | 0.1349 | 0.0652 | 0.0338 | 0.0328 | 0.0328 | 0.2025 | 0.0675 |
| C4 | 0.0725 | 0.0823 | 0.0456 | 0.0456 | 0.0805 | 0.0923 | 0.1972 |
| C5 | 0.1349 | 0.0619 | 0.0341 | 0.0341 | 0.0355 | 0.1693 | 0.0991 |
| 续表 | | | | | | | |
|  | 指标权重 |  |  |  |  |  |  |
| C1 | 0.1598 | 0.0239 | 0.0655 | 0.0141 | 0.1307 | 0.0149 | 0.1598 |
| C2 | 0.0666 | 0.0666 | 0.0666 | 0.0666 | 0.0666 | 0.0589 | 0.0666 |
| C3 | 0.2025 | 0.0361 | 0.0422 | 0.0452 | 0.0698 | 0.1696 | 0.2025 |
| C4 | 0.1752 | 0.0472 | 0.0412 | 0.0463 | 0.0897 | 0.0568 | 0.1752 |
| C5 | 0.2063 | 0.0344 | 0.0510 | 0.0353 | 0.0706 | 0.1683 | 0.2063 |

记的指标权重为 ()；在类别下，特征的权重为。可那么可根据下面的算式求出特征前的权重

又考虑到的实际含义分别是是首次发球失误、双误、非受迫性失误，会对势头产生消极的影响，因此认为三个指标与势头是负相关的。得到的最终权重如下表所示

表 12个特征的最终权重

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 权重 | 值 |  | 权重 | 值 |  | 权重 | 值 |  | 权重 | 值 |
|  | -0.0506 |  |  | -0.1202 |  |  | 0.1662 |  |  | 0.0273 |
|  | 0.0645 |  |  | -0.0729 |  |  | 0.0318 |  |  | 0.1067 |
|  | 0.1452 |  |  | 0.0933 |  |  | 0.0587 |  |  | 0.0627 |

从表中可以看出，的绝对值最大，说明（破发点）对原始得分的影响最显著；的绝对值最小，说明（得分差异比值）对原始得分的影响最微弱。

## 4.3 选择平滑系数

根据上述分析以及实际情况，我们认为在t时刻下，若存在休息时间，对的影响不大。因此，初步选择前的系数为0.35，即模型的平滑系数为0.65。得到势头的EWMA模型为

然后根据公式()求出球员1和球员2的势头。进一步地，将我们得到的数据与比赛真实胜负、选手表现等多方面进行比对，以此验证我们选择的系数是否恰当。

下面展示在“2023-wimbledon-1301”这场比赛中，两位球员的势头表现。

图 球员1的势头

这是平滑指数在0.65下的选手1的势头随比赛流程变化的折线图，经过和比赛真实胜负、选手表现指标等多方面进行比对，我们发现在该模型下选手的势头与其真实表现最为吻合。

图 球员2的势头

这是平滑指数在0.65下的选手2势头随比赛流程变化的折线图，经过多方面比对，该模型下选手的势头与其真实表现最为吻合。

所以我们选择0.65为最终模型的平滑指数。

## 4.4 可视化比赛流程

4.4.1 确定表现更好的球员

为了比较两位球员的表现，我们引入了“相减势头”

若大于0，则球员1表现更好；若小于0，则球员2表现更好.下面展示在“2023-wimbledon-1301”这场比赛中，两位球员的“相减势头”。

图 两位球员的相减势头

从图中看出，两位球员的相减势头变化波动大、频次高，并且在0点附近均匀分布，说明比赛中每点的比分争夺激烈，双方的水平总体表现相差不大。

4.4.2 球员的具体表现

至于表现更好的球员到底表现有多好，可以用公式()求出该球员在t时刻下的势头，以关注该球员本身的表现。比如，图()和图()可视化了“2023-wimbledon-1301”这场比赛两位球员各自的具体表现。

# 5 Task 2 探究势头的作用

这一部分，我们主要考虑了势头与比赛结果的线性关系和因果关系。为了研究的方便，我们用“相减势头”反映“势头”，用“得分”反映“比赛结果”。首先我们对两者进行了ADF检验，然后进行了交叉相关分析和Granger因果检验，用于分别探究它们之间是否存在线性关系或者因果关系。另外，我们也对“相减势头”进行了自相关检验。

## 5.1 探究相减势头的时间序列是否随机

在考虑“相减势头”与比赛结果的关系之前，我们想先把目光聚焦到势头本身，考察它的时间序列是否是随机的。在进行自相关检验之前，我们通常需要检验时间序列的平稳性.

5.1.1 平稳性检验

1.模型的检验

Augmented Dickey-Fuller Test（ADF检验）是最常见的单位根检验之一，用于检验一个时间序列是否具有单位根，从而判断其平稳性。具体模型如下，

设原假设为存在单位根，时间序列是非平稳的；设备择假设不存在单位根，时间序列是平稳的，即

假设时间序列为

其中，是差分操作符， 是时间序列观测值，是时间趋势项，是截距，是白噪声误差。

计算检验统计量的p值，若p值大于0.05，则无法拒绝原假设，说明时间序列是非平稳的，若p值小于0.05，则拒绝原假设，说明时间序列是平稳的。

2.模型的求解

根据软件我们得到了在1%、5%和10%的显著性水平下的ADF检验结果，

表 相减势头的ADF检验结果

|  |  |
| --- | --- |
| 相减势头 | |
| ADF Statistic | p-value |
| -10.316684188792774 | 3.0841757502686896 |

从表中容易得知p-value远小于0.05，表明在1%、5%和10%的显著性水平下，我们都可以拒绝存在单位根的原假设，从而认为数据是平稳的。

5.1.2 自相关性的检验

1.模型的建立

ACF和PACF是用于时间序列分析的两种常见的自相关性检验方法，旨在检查时间序列中不同滞后阶数之间的自相关关系。下面是相关公式：

1. ACF

对于滞后阶数，计算时间序列和之间的自相关系数为

1. PACF

对于滞后阶数，计算时间序列和之间的偏自相关系数，

其中，是对进行回归得到的预测值。

2.模型的求解

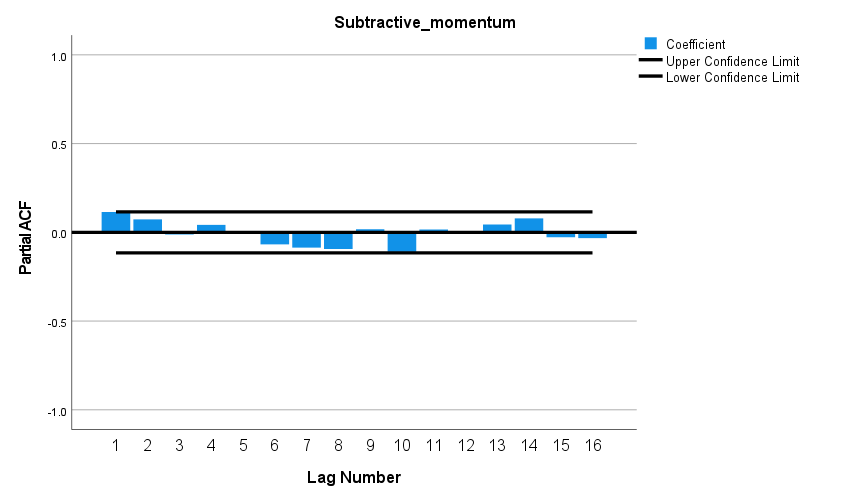
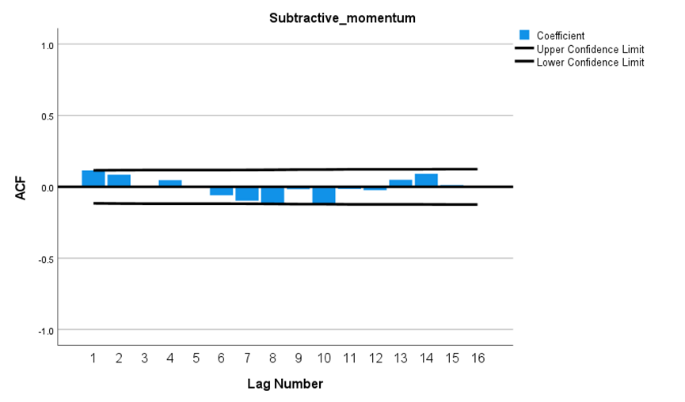


图 残差的ACF和PACF图

从图中可以看出，所有滞后阶数的自相关系数和偏自相关系数均和0没有显著的差异，表明残差是一个白噪声序列。

## 5.2 讨论势头与比赛结果的关系

对于比赛结果，我们考虑用“累计得分”或者“得分”来刻画，因此问题转化为“讨论相减势头与累计得分或者得分的关系”。进一步地，我们主要想讨论两者的线性关系以及因果关系。但在此之前，通常要求两者的时间序列是平稳的，所以进行平稳性检验。

5.2.1 累计得分、得分时间序列的平稳性检验

在前面我们已经检验了相减势头的平稳性，因此这部分只对累计得分和得分做检验。下面展示结果。

表 累计得分和得分的ADF检验结果

|  |  |  |  |
| --- | --- | --- | --- |
| 累计得分 | | 得分 | |
| ADF Statistic | p-value | ADF Statistic | p-value |
| 0.763834912212 | 0.991029388897 | -14.07951731660 | 2.84517110749 |

从表中容易得知，累计得分的p-value明显大于0.05，表明在1%、5%和10%的显著性水平下，我们都无法拒绝存在单位根的原假设，从而认为数据是不平稳的。但是得分的p-value远小于0.05，表明在1%、5%和10%的显著性水平下，我们都可以拒绝原假设，因此数据是平稳的。

总而言之，相减势头和point\_victor这两列数据都是平稳的，可以直接用于进一步的建模工作。

5.2.2 探究相减势头与得分的线性关系

1.模型的建立

交叉相关分析（Cross-Correlation Analysis）用于测量两个时间序列在不同时间点上的相关性。其原理是通过计算不同时间滞后下的相关系数来描述两个序列之间的关系。以下是交叉相关分析的基本原理和数学公式：

针对相同样本长度N的时间序列和，用表示滞后时间，为正表示滞后于，为负表示滞后于。在滞后时间处的交叉相关系数为

一般来说，如果在某个特定的滞后时刻，相关系数显著不为零（即置信区间不包含零），这意味着在该滞后时刻两个时间序列有显著的相关性。

2．模型的求解

通过计算不同滞后时间下的相关系数，可以绘制出相减势头与得分的交叉相关函数图。

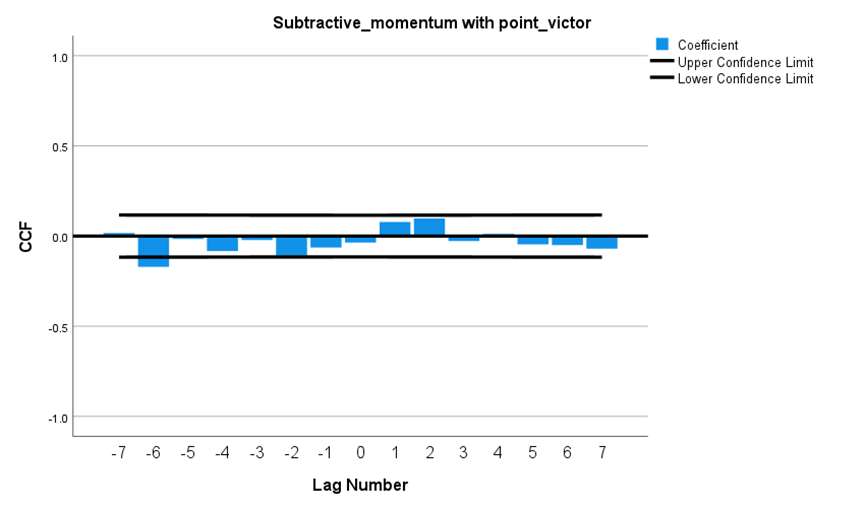


图 相减势头与得分的交叉相关函数图

从图中我们可以看到，所有的滞后点的置信区间都包含了零，说明在观测的滞后范围内，相减势头与得分没有显著的线性关系。

这样的结果可能表明，两个时间序列是独立的，或者它们之间的关系可能是非线性的，或者相关性可能存在于更高阶的动态关系中。

5.2.3 检验相减势头与得分的因果关系

1.模型的建立

Granger causality test是一种用于检验时间序列数据之间因果关系的统计方法，可以研究变量之间的因果关系。以下是因果检验的模型：

设原假设为时间序列变量X不是引起时序变量Y变化的原因，备择假设为时间序列变量X是引起时序变量Y变化的原因，即

Null Model为

Alternative Model为

然后使用最小二乘法对上述模型的参数进行估计，再利用这两个回归模型的残差平方和和得到检验统计量和p值。

如果检验统计量显著或p值小于0.05，则拒绝原假设，说明存在格兰杰因果关系。否则，无法拒绝原假设，说明未发现显著的因果关系。

2.模型的求解

求解后，我们得到了相减势头与得分的因果检验结果，

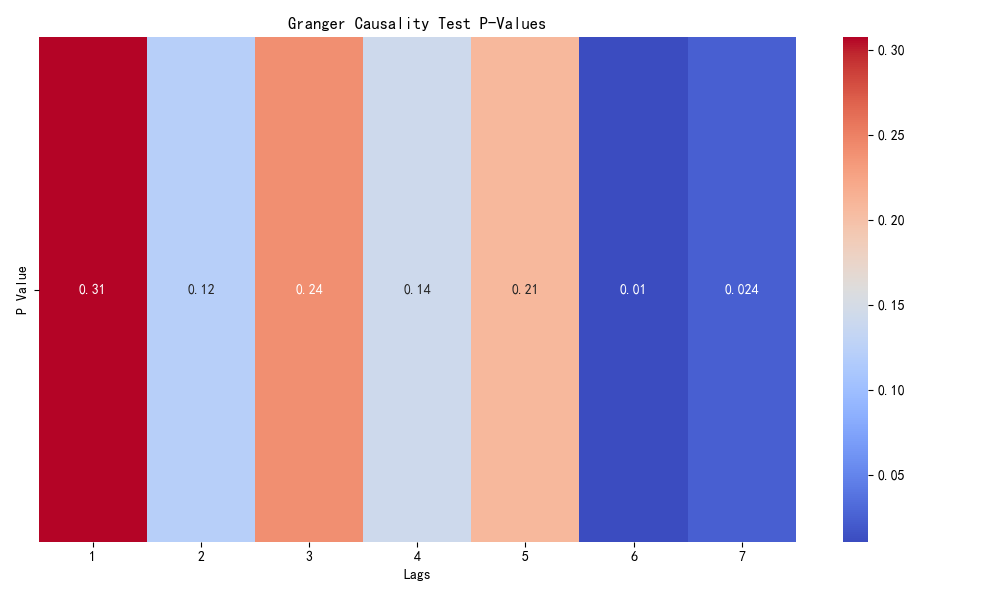


图 相减势头与得分的因果检验结果

由图可知，滞后期为6及之后的p值小于0.05，因此拒绝原假设，说明相减势头是得分的格兰杰原因；并且在滞后期较长时，p值非常小，说明显著性更强。

## 5.3 小结

势头在观测的滞后范围内与得分之间没有显著的线性关系。但是在较长的时间尺度上，势头对得分的影响更为显著。说明势头不仅仅是短期内的表现，而是在比赛历程中逐渐积累的趋势，对比赛结果有着长期的、显著的影响。

# 6 Task 3 探究比赛流程的转变

【来一个总起的段落：在这一部分，我们-----】

## 6.1 预测波动

我们将比赛中的波动用“累计相减势头”来刻画，以此从宏观上捕捉比赛的变化。因为累计的相减势头在时间上相互关联，因此我们选择时间序列模型进行预测。

6.1.1 作时序图

我们在时间序列模型中，将时间定义为了“小时：分钟”，也就是每一次的发球时间为每一分钟。这放宽了发球的具体时间，在一定程度上方便了时间序列模型的构造。下面展示累计相减势头的时间序列图。

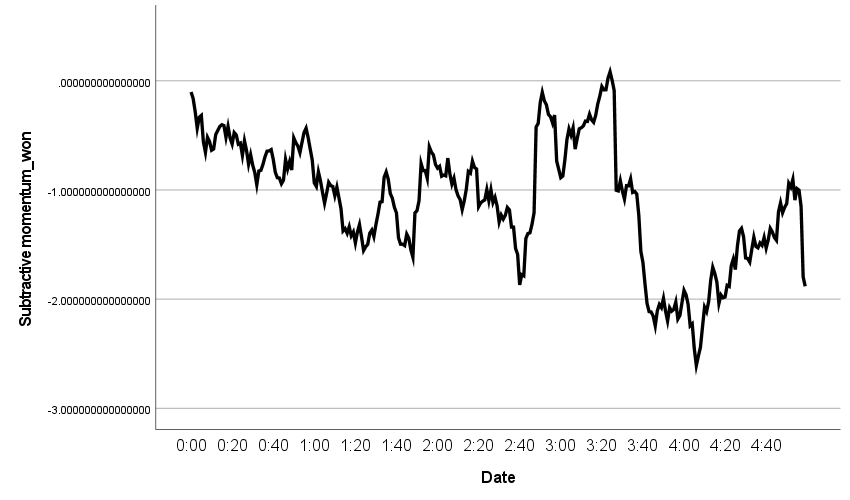


图 累计相减势头的时间序列图

从图中可以看出，随着时间变化，相减势头总体的波动较大，但在某些时间段内波动较小，因此可使用加法分解模型。另外，可以发现数据一般在整点时发生较大突变，因此猜测可能每60min为一周期。

6.1.2 时间序列分解

对时间序列进行分解，可以发现周期为60min。同时也可以得到对应的季节性因子，再将季节性因子绘制在时间序列图上以直观地观察季节性变化。

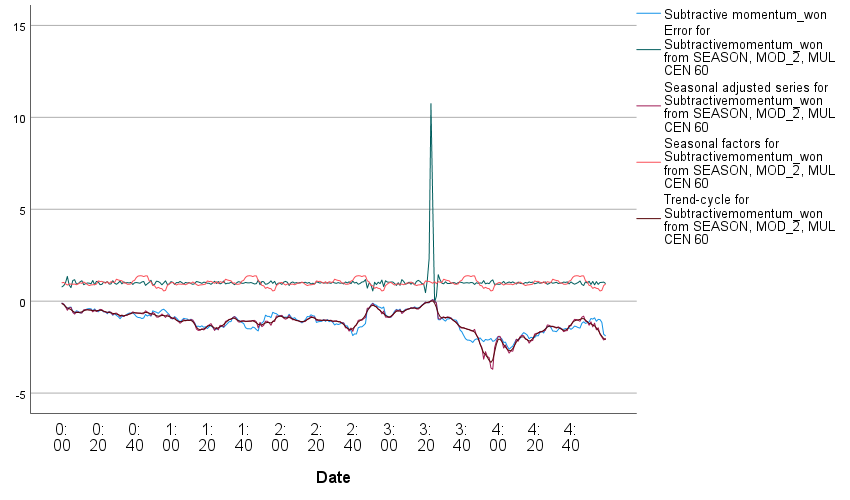


图 带有季节性因子的时间序列图

由图可知，橙色曲线有明显的周期性波动，说明比赛总体势头是存在一定周期性的。绿色曲线对应的值较小，说明模型拟合较好，但有一段有明显突变，这可能是数据中有异常值或者比赛出现了模型没有考虑到的重大事件影响了比赛形式导致的。红色曲线与蓝色曲线拟合较好，说明模型拟合优度高。

6.1.3 模型的建立、求解与检验

利用SPSS软件专家建模得到模型类型为Winters' Additive，具体公式如下

其中，为当期，为预测步长，为周期长度（这里取60），为时刻的预估水平，为水平的平滑参数，为趋势的平滑参数，为季节的平滑参数，为第期的预测值。

同时，也得到了对应的系数，分别是，，，将它们代入公式()即可。

接着我们对模型进行检验。因为时间序列模型通常假设残差是白噪音，因此我们选择白噪声残差检验。

具体的，我们采用Ljung-Box检验。这是一种常用的方法，用于检验残差序列是否存在自相关性。

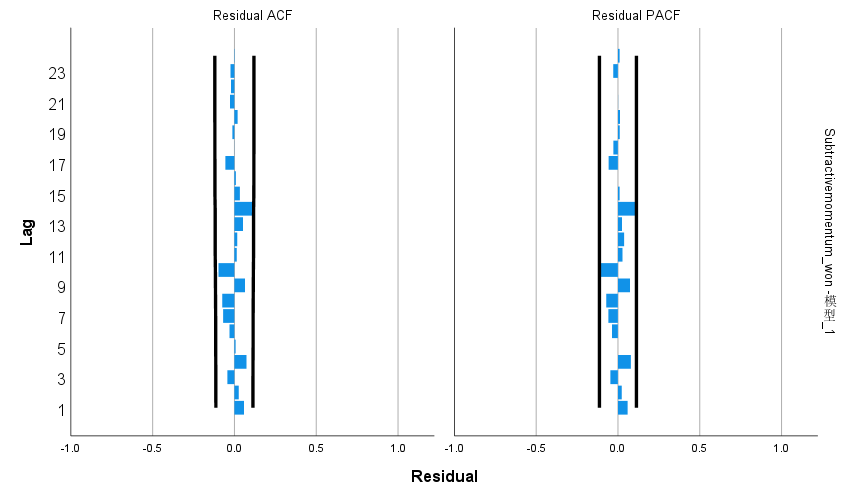


图 白噪声残差检验

表 模型数据

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number of Predictors | Stationary R-squared | Statistics | DF | Sig. | Number of Outliers |
| 0 | 0.604 | 17.674 | 15 | 0.280 | 0 |

从残差的ACF和PACF图形中可以看出，所有滞后阶数的自相关系数和偏自相关系数均和0没有显著的差异；从表中可以看出平稳 R 方为0.604，数值较大，说明模型很好地适应了平稳化后的时间序列。而且，对残差进行Q检验得到的p值为0.280，大于0.05，无法拒绝原假设，因此认为残差就是白噪声序列，说明加法模型能够很好的识别本题中的销量数据。

6.1.4 预测以及对结果的检验

## 6.2 寻找与波动最相关的因素

6.2.1 平稳性检验

相关的方法前文已经介绍过了（见公式（）），下面仅展示结果。

表 跳过

表 “累计相减势头”与其他因素的ADF检验结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 指标 | p-value | 指标 | p-value | 指标 | p-value |
| double\_fault | 5.441 | y81 | 2.938 | M(t) player2 | 4.261 |
| winner\_shot\_type | 2.125 | y82 | 4.067 | Subtractive momentum\_won  （一阶差分后） | 2.829 |
| Serve | 3.901 | y9 | 1.494 |  |  |
| serve\_no | 3.613 | Y10 | 0.026 |  |  |
| unf\_err | 4.038 | ace | 3.152 |  |  |
| net\_pt | 3.183 | score\_rate | 3.072 |  |  |
| break\_pt | 1.145 | M(t) player1 | 2.709 |  |  |

从表中容易得知所有指标的p-value均小于0.05，表明在1%、5%和10%的显著性水平下，数据是平稳的。因此可以用于下一步的工作。

6.2.2 交叉相关分析

为了探究“累计相减势头”与其他因素的时间序列在不同时间点上的相关性，我们采用交叉相关分析。需要注意的是，这里选择的是一阶差分后的“累计相减势头”时间序列数据，因为它是平稳的。

此分析的方法已经介绍过了（见公式（）），下面具体分析几个具有代表性的结果，其余的仅给出结论。

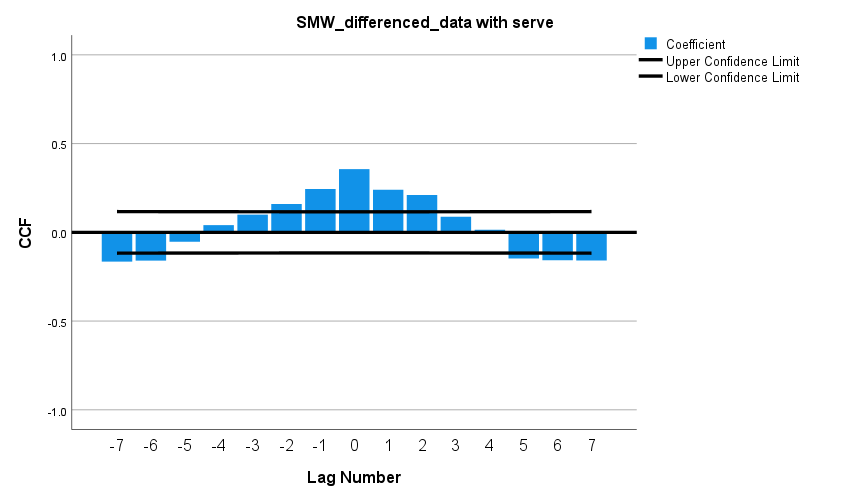
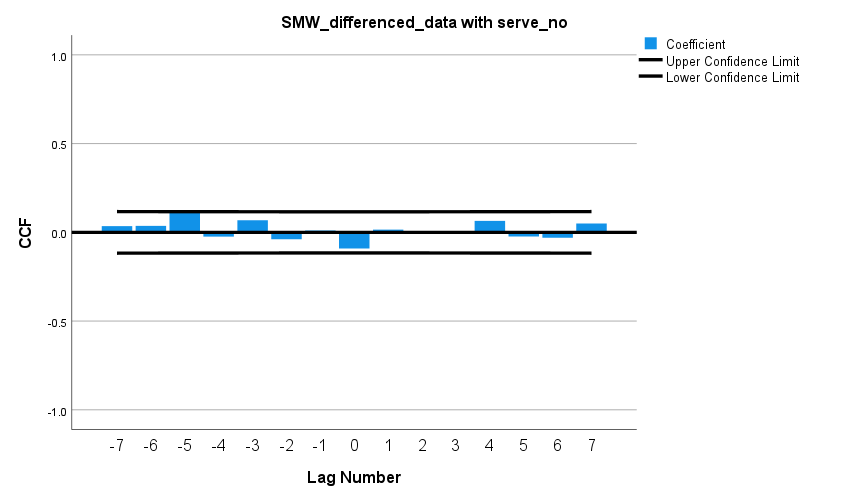
1.对个别的分析

图 serve 图 发球失误

由左图可以看出，在大多数滞后期上，serve和比赛累计相减势头有着较为明显的线性关系且呈周期性分布，这可能是受到发球方周期性变换的影响。部分呈现不明显线性关系，可能是因为统计时受到交换发球的影响，将发球优势抵消了。而在滞后0阶表现出较强的线性关系，可能是因为球员存在一直发球的优势。

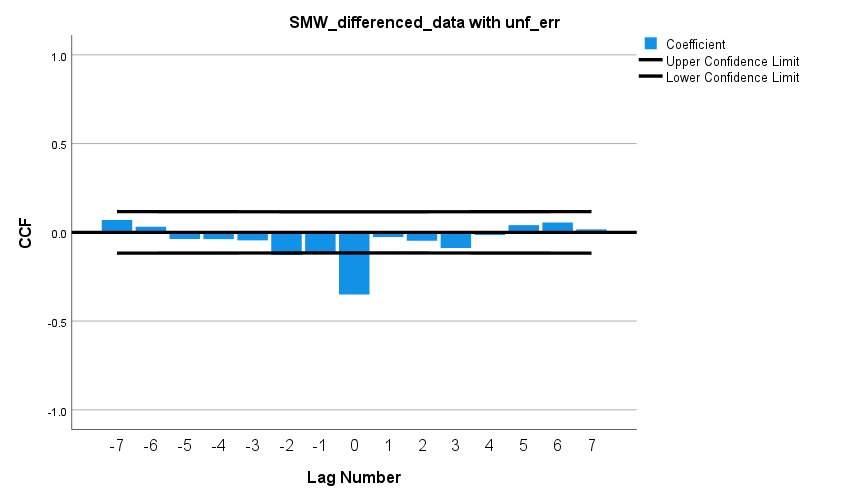
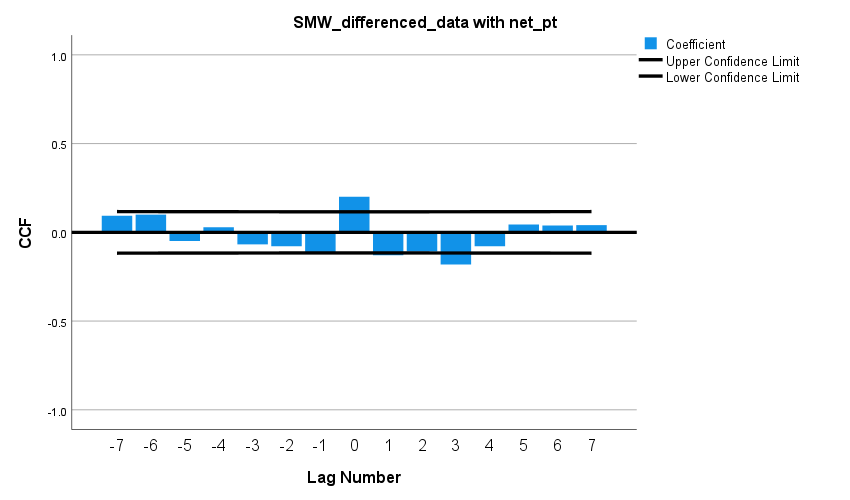
由右图可以看出，发球失误对累计势头的改变没有明显线性关系，这可能是因为没有将发球失误数据与发球球员结合起来。

图 Unf\_err 图 Net put

由左图可以看出，在滞后期为0阶时，Unf\_err和比赛累计相减势头有明显的负相关线性关系，这可能受到了专业比赛中的非受迫性失误的即时性和稀有性的影响。

由右图可以看出，在滞后期为0阶和3阶时，Net put和比赛累计相减势头有一定的线性关系，且一个是正相关一个是负相关，这可能是我们选用成功上网数据不知道是哪位球员成功上网导致的。

2.结论

根据结果，我们将各项指标按照与累计相减势头的关系强弱分类：

* 有明显线性关系：Serve，Unf\_err，Unf\_err（滞后期为0阶时）
* 关系较弱或不明显：Double\_fault，发球失误，Winner\_shot\_type，Net put，Break\_pt，球员的体能指标，
* 无明显线性关系：y9,y10,ace,得分率

6.2.3 格兰杰因果检验

下面我们进一步想知道“累计相减势头”与其他因素是否存在因果关系，因此采用格兰杰因果检验。相关的方法已经介绍过了（见公式（）），下面展示检验结果。

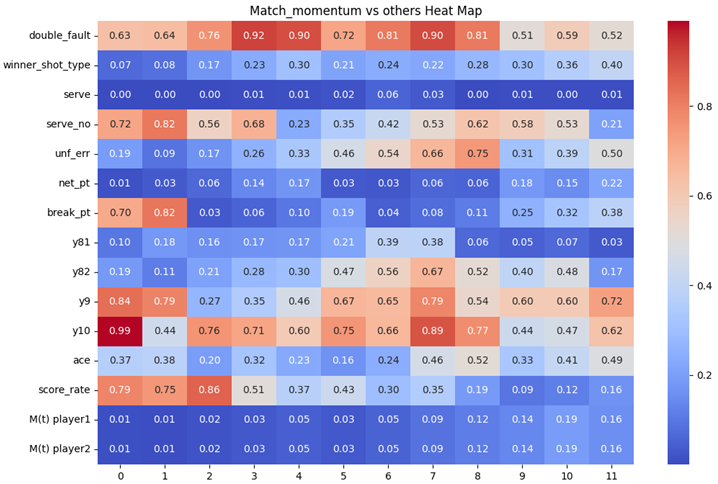


图 “累计相减势头”与其他因素的因果检验

从图中我们可以得出以下信息：

* winner-shot-type在滞后期1~2时p值小于0.1，较为显著，说明其与比赛的波动有一定的因果关系。
* serve在各个滞后期上p值均小于0.05，非常显著，说明其和比赛的波动有着明显的因果关系。
* M(t)player1和M(t)player1的p值随着滞后期的变大而变大，且在滞后期为0~6时p值小于0.05显著，说明其和比赛的波动有着明显的因果关系。
* net-put在滞后期为0~1、5~6时p值小于0.05，非常显著，说明其和比赛的波动有着明显的因果关系。
* Y81（每拍跑动距离）在滞后期为8、10 时p值小于0.1，较为显著，说明其与比赛的波动有一定的因果关系；在滞后期为9、11时p值小于0.05，非常显著，说明其和比赛的波动有着明显的因果关系。

6.2.4 小结

根据上述分析，总体而言，Serve、Unf\_err、M(t)player1、M(t)player2、Net\_put是与比赛波动最相关的因素，而其他因素的关系相对较弱或不明显。

## 6.3 提供建议

考虑到过去比赛“势头”波动的差异，我们根据模型研究提出以下建议：

* 当上一场比赛的累计总势头高于对手并持续上升时，球员应该在下一场比赛中加快比赛节奏，以继续拉大比分差距，巩固优势地位。
* 当上一场比赛的势头持平时，要特别注意每一个点的发挥，积极展现出精彩的操作，如ace等。同时，要避免非迫性失误，如sever no等，争取达到更有利的地位。
* 当上一场比赛的势头处于劣势时，应该放缓比赛节奏，同时等待时机进行反击，以扭转劣势。
* 在上一场比赛势头出于优势但优势地位有下降趋势时，下一场比赛需保持警惕，防范可能的反击。
* 当上一场比赛的势头虽然处于劣势地位但有上升趋势时，下一场比赛应更加积极，争夺势头，全力争取扭转局势。

# 7 Task 4+解决的问题

【来一个总起的段落：在这一部分，我们-----】

## 7.1 aaaaaaaaaaaa

## 7.2 aaaaaaaaaaaa

# 8 备忘录

# 9 Sensitivity Analysis（敏感性分析）

# 10 Model Evaluation and Further Discussion

模型一：优点利用球员获得杠杆率的指数加权移动平均模型 简单、灵活，特别适合于处理有趋势和季节性的时间序列数据。能很好的定量刻画球员的势头。

缺点：选择合适的平滑指数值有时候比较主观，可能需要通过多次试错或优化方法来确定最佳值。

模型二：优点：对相减势头和point victor 的序列进行了一系列检验如ADF检验 交叉性相关检验 Granger因果检验 自相关性检验等，证明了相减势头和比赛结果有明显的因果关系

缺点：没有找到势头和比赛结果的具体什么关系如线性还是平方还是指数还是几个函数的组合。

模型三：优点：利用历史数据测算出 累计的相减势头序列进行了时间序列模型预测，能很好的描绘整场比赛的流程和比赛总势头，且预测结果在一段时间内的形式较好。通过对累计相减势头与其他指标进行的交叉相关性检验和因果检验能很好的看出哪些具体指标是波动产生的重要指标，及这些指标是怎样影响整场比赛的局势的。对比赛策略的调整具有很好的参考价值。

缺点：该模型无法自动实时更新比赛的历史数据随着比赛时间的推移预测准确度可能会下降，且由于前期比赛开始时历史数据较少模型预测结果可能也不太精确。

## 10.1 Strength（优点）

## 10.2 Weaknesses（缺点）

# References【单独一页】

【期刊名用斜体】

[1]

[2]

[3]

[4]