电商零售商家需求预测及库存优化问题研究

摘要

线上购物是近些年来兴起的新购物浪潮之一，电商平台及其零售商家作为线上购物的主要支持方，面对不同的市场浪潮，基于历史数据通过算法得到历史一般规律，对下一阶段的需求量进行预测，进而支持库存优化等后续相关问题的决策，是一项具有重要应用意义的工作。

针对问题一，我们对数据整理得到了 1996 个**时间序列**，并通过**回归随机森林模型**对时间序列进行了均值回归，并引入均方差（MSE）和 Durbin-Watson（DW）检验来衡量模型的预测准确率。另外，我们建立了**历史波动随机变量模型**来衡量时间序列的波动规律，作为**库存优化**等后续决策的另一方面支持。对于问题一的分类问题，我们引入**动态时间规整模型**（DTW）来衡量时间序列的相似程度，在 DTW 的基础上设计了**基于 DTW 的改进 K 均值聚类模型**，并根据附件 2-4 中的语义标签建立了**相似度检验模型**，用来衡量聚类分析的聚类效果，并指导聚类模型的超参数调整，最终得到分为 15 类的时间序列聚类簇，其同一类别在需求上的特征最为相似。

针对问题二，我们采取**残缺时间序列补全模型**和**迁移学习模型**相结合的方法，对附件 5 的时间序列进行了均值回归。我们在问题一的基础上采用 DTW 模型和相似度检验模型，以新维度的时间序列为聚类中心，依次对问题一中的时间序列进行**归类**。采用归类的时间序列对残缺时间序列进行补全，并采用问题一中已经经过大量训练、结构优化和参数迭代的回归随机森林模型，对**补全时间序列**进行均值回归，得到预测结果。

针对问题三，我们考虑到出货量的**时间序列波动趋势**也类似于一种信号，在较大范围内拥有周期性变化的特征，其峰值的到达常常与外界因素有关，如“大型促销”等。我们借鉴**太阳黑子**周期性变化规律中的研究方法，引入**电商波动相对指数**，并对**现实情况**进行了相关调查，建立了**叠加电商波动相对指数的均值回归模型**，采用该模型对附件 6 的时间序列维度进行回归分析，得到包含**波动趋势**的预测结果，以支持大型促销等特殊市场浪潮的优化决策。

关键词： 回归随机森林　相似度检验　历史波动随机变量　基于 DTW 的改进 K 均值聚类　电商波动相对指数

目录

一、 问题重述

1.1 背景

电商平台存在着上千个商家，他们会将商品货物放在电商配套的仓库，电商平台会对这些货物进行统一管理。大数据智能驱动的供应链研究可以通过分析历史一段时间每种“商家”、“仓库”、“商品”维度的出货量数据，预测未来一段时间的时间序列的走势，显著降低库存成本，同时保证商品的按时履约。同时，企业会根据数据的历史情况，分析出需求量序列的数理特征，对相似的需求量序列进行归类，并根据归类结果做到更加精准的预测。然而，在实际的电商供应链预测任务中，常常会出现多种复杂情境。如部分商品销售时间过短、仓库存在新增或切换等情况，导致该预测维度下历史数据过少；另外，部分大型促销期间货量的陡增并由此带来的不规律性，也给需求量的精准预测带来了不小的难度。此时便需要通过算法得到历史一般规律，找出相似的历史情况，从而实现精准预测。

1.2 问题提出

题目要求按照所给的需求预测和库存优化等电商供应链场景，结合附件 1 至 6 的数据，完成以下问题：

问题一：使用附件 1-4 中的数据，预测出各商家在个仓库的商品 2023-05-16 至 2023-05-30 的需求量，请将预测结果填写在结果表 1，并对你们模型的预测性能进行评价。另外请讨论：根据数据分析及建模过程，这些由商家、仓库、商品形成的时间序列如何分类，时同一类别在需求上的特征最为相似？

问题二：现有一些新出现的商家 + 仓库 + 商品维度，导致这种情况出现的原因可能是新上市的商品，或是改变了某些商品所存放的仓库。请讨论这些新出现的预测维度如何通过历史附件 1 中的数据进行参考，找到相似序列并完成这些维度在 2023-05-16 至 2023-05-30 的预测值。

问题三：每年 6 月会出现规律性的大型促销，为需求量的精准预测以及履约带来了很大的挑战。附件 6 给出了附件 1 对应的商家 + 仓库 + 商品维度在去年双十一期间的需求量数据，请参考这些数据，给出 2023-06-01 至 2023-06-20 的预测值。

二、 问题分析

2.1 数据预处理

本题附件 1 给出的数据共 331336 行，附件 5 给出的数据共 7429 行，附件 6 给出的数据共 21355 行，数据量较为庞大且排列顺序并不按时间戳进行，为使后续工作顺序进行，我们编写 python 程序对附件数据进行了预处理，进行了数据清洗，排除了重复和

异常项，按照时间戳对数据进行重新排列，并将商家、商品和仓库的字符串名称进行处理，便于后续的数据提取和转化工作。

2.2 问题一分析

本题目在于在庞大的数据量中提取每个维度的全部时间序列，并针对电商平台出货量这种特殊类型的时间序列寻找一种性能较好的回归模型，对此我们采取回归随机森林进行均值回归，同时建立历史波动随机变量，用于体现时间序列的波动规律。此外，题目要求根据数据分析和建模的过程，对时间序列数据集进行分类，使其在需求上的特征最为相似，对此我们采取基于 DTW 距离的改进 K 均值聚类分析，对 1996 个时间序列进行聚类，并根据附件 2-4 建立相似度检验模型，对聚类结果进行进一步检验。

2.3 问题二分析

本题目在于有一些新出现的商家 + 仓库 + 商品维度，其时间序列长度较短，无法支撑回归分析。我们考虑通过附件 1 的数据进行参考，基于第一问的 DTW 模型和相似度检验模型，以新维度的时间序列为聚类簇中心，对附件 1 的时间序列进行归类，并用归类结果补全新维度的时间序列。最后，我们采取迁移学习，利用问题一中以及训练成熟的回归随机森林对问题二的补全时间序列进行回归分析。

2.4 问题三分析

本题目考虑每年 6 月出现的规律性的大型促销会给商品需求量带来较大波动，希望参考去年双十一期间的需求量数据对 2023 年 6 月份的 20 天的需求量进行预测。我们考虑到出货量的时间序列波动趋势类似于一种信号，在较大范围内拥有周期性变化的特征，其峰值的到达常常与外界因素有关，如“大型促销”等。所以我们参考通过太阳黑子研究太阳周期性活动的研究过程，对规律性的大型促销时间段引入电商波动相对指数，进行叠加电商波动相对指数的均值回归。同时，我们结合对现实情况的调查，对 6 月 18 日前后的需求量数据进行特殊处理，使其更符合现实需求。

三、 模型假设

为了便于考虑问题，我们在不影响准确性的前提下，做出以下假设：

1. 假设时间序列维度的商家、商品和仓库相应的的语义标签的相似与时间序列数理特征的相似具有内在关联。
2. 将数据作为首要的、决定性的建模依据，回归分析过程中不考虑超出数据描述范围的异常情况的影响。
3. 假设大型促销的时间范围较大，从 6 月初就已经开始。

四、 符号说明

符号	说明
$Q(T)$	出货量时间序列
$Q_{pre}^{rf}(T)$	随机森林回归得到的预测时间序列
$Q_{pre}^{cr}(T)$	含波动修正的预测时间序列
$D_{pre}(T)$	波动修正的抽样序列
K	分类总数
K_k	第 k 个分类簇中心
S_k	第 k 个分类/时间序列集合
Q_i^k	第 k 分类的第 i 个元素/时间序列
X, Y, Z	时间序列的三种一级语义标签
A	时间序列语义标签对应的 8 维数阵
A_i	A 数阵的元素，单位矢量，对应一种语义标签
D_{dtw}	时间序列 DTW 距离矩阵
R_{sim}	相似度检验函数的相关矩阵序列
R^m R_{ij}^m	相似度检验函数的相关矩阵及元素
s_i	商品特征类似程度，用于产生 R_{sim} 的矩阵元

表 **1**　符号三线表说明

五、 模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 预测策略的确定

在本次整个数学建模问题中，我们考虑到主要任务是针对库存方和平台方的出货量进行预测，其根本目的是通过预测为库存优化和后续一系列决策提供支持，单个时间序列在仅有历史出货量数据支持的前提下，精确到每一天的准确预测在理论上的可行性空间较小，在决策支持上给予的数据支撑单薄。

我们认为，时间序列的预测应当从两个方面考虑，一是每种时间序列过滤大幅度波动后的均值回归，这反映了商品在无较大外部因素扰动下的平均需求量，这是考虑库存优化时的决策基准；二是每种时间序列的波动规律，包括波动幅度，波动周期等，这反映了商品受外界多方面因素影响时偏离回归均值的归类，以便决策者在面对市场规律浪潮时为库存优化等一系列决策做出更加充分的准备，同时预留出适当的缓冲空间和周转余地。综合上述两种要素，决策者可以在多方面因素的影响下做出更优的决策。而我们后续三个问题的建模与求解，均基于上述思想进行，在填写预测结果表时，我们考虑到不同情境下的决策重点，从而给出更有助于优化决策的预测值数据集。

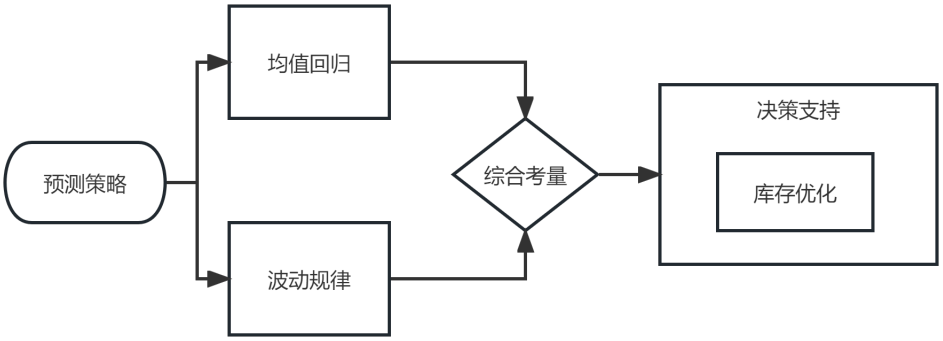


图 **1**　预测策略导图

5.1.2 时间序列的确定

对于给出的数据集附件 1-商家历史出货量表，我们编写 python 程序对其进行了数据清洗，排除其重复数据和异常数据，然后筛选出所有的商品 + 商家 + 仓库维度，并将其对应的出货量按照时间戳进行排序，得到了 1996 个时间序列 $Q(T)$, 每个 $Q(T)$ 长 166 天，其第 i 个维度组合对应的 $Q(T)$ 为：

$$[X, Y, Z]^i : Q(T) = [Q(1), Q(2), \cdots, Q(n_i)]^T \tag{1}$$

5.1.3 随机森林

随机森林属于集成学习算法的一种，其核心思想是集成多个弱分类器，从而实现一个预测效果更好的集成分类器。随机森林可以同时胜任回归和分类两大任务。随机森林基于决策树构建而成，主要思想是通过组合多个决策树来提高模型的性能和稳定性。

随机森林的关键特点和工作原理：

1. 随机选择特征：在构建每棵决策树的过程中，随机选择一部分特征进行训练，而不是使用所有的特征。这有助于减少过拟合的风险，并增加模型的多样性。
2. 随机采样数据：对于每棵树的训练数据，随机抽取部分样本进行训练，这被称为自助采样（bootstrap sampling）。这意味着某些样本可能在多棵树的训练中被重复采样，而某些样本可能从未被选中。
3. 多树组合：随机森林通常包含多棵决策树，这些树独立构建，并且没有共享信息。在分类问题中，每棵树投票决定最终的类别；在回归问题中，每棵树的预测结果取平均值。
4. 随机性和多样性：随机森林的随机性来自于特征选择和数据采样，这有助于防止过拟合。多棵树的组合提供了模型的稳定性和鲁棒性，使其适用于各种问题。

其基本结构示意图如图2所示。随机森林算法在处理大规模时间序列的优点在于：

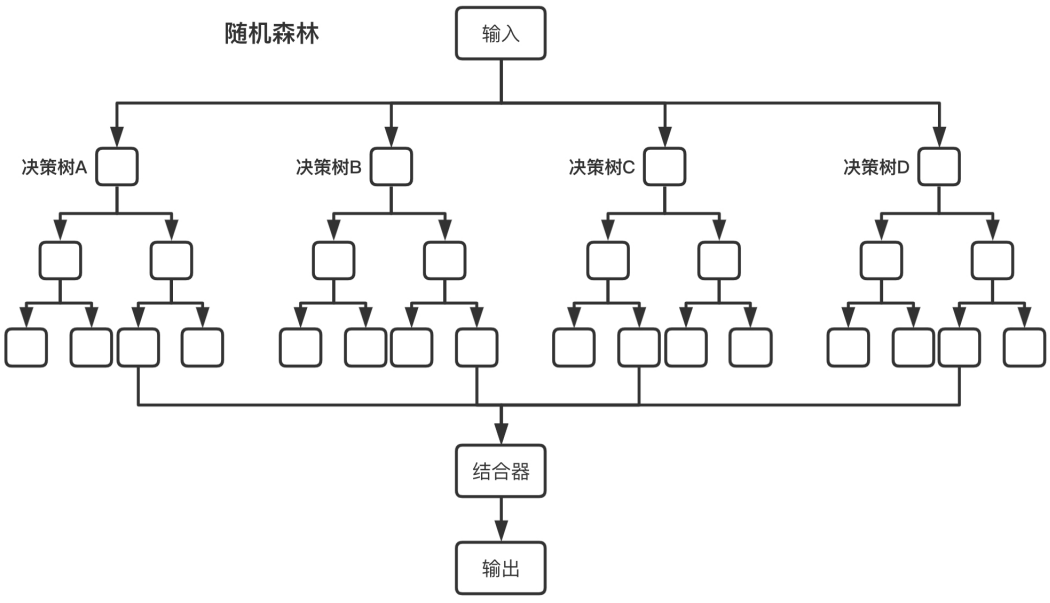


图 **2**　随机森林基本结构

1. 在建造森林时，可以在内部对于一般化后的误差产生不偏差的估计
2. 可以有效处理大量的输入变量，学习过程迅速
3. 对于不平衡的时间序列资料集来说，可以平衡误差

4. 包含一个好方法可以估计丢失的资料，并且如果有很大一部分的资料丢失，仍可以维持准确度
5. 可被延伸应用在未经标记的时间序列上

5.1.4 回归随机森林

基于随机森林的设计思想，我们设计出基于时间序列的回归随机森林。在回归随机森林中，多个决策树组合起来进行预测，最终输出是这些树的预测结果的平均值。与单一决策树相比，回归随机森林通常能够提供更稳健的预测，减少过拟合的可能性，并对数据中的噪声更具鲁棒性。其过程示意图如图：

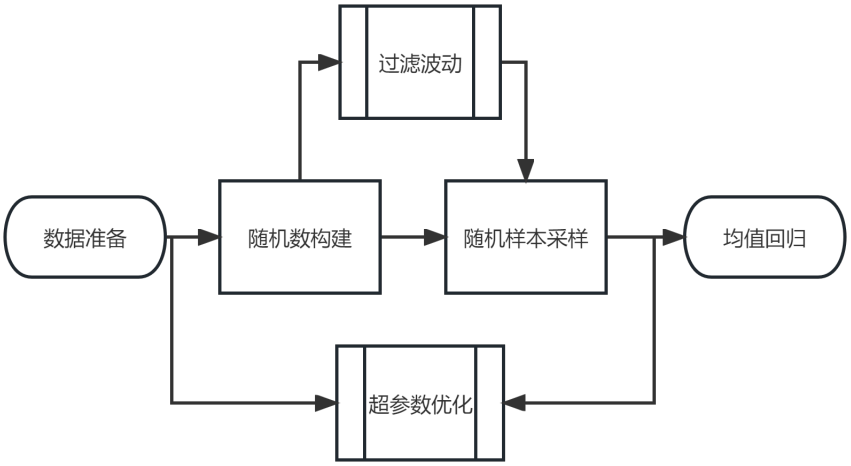


图 **3**　回归随机森林流程图

5.1.5 预测准确率检验

对于验证集数据的检验方法，我们认为加权平均绝对百分比误差 (wmape) 在这种模型中的参考价值已不大，因为我们过滤时间序列中的波动规律，求得一般性的均值回归的过程中，峰值处的较大误差会使整个 wmape 在数值上失真。所以，我们考虑均方误差 (MSE) 和 Durbin-Watson(DW) 检验相结合的方法构建预测准确率检验模型：

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n} \tag{2}$$

$$DW = \frac{\sum_{t=2}^{n} (\epsilon_t - \epsilon_{t-1})^2}{\sum_{t=1}^{n} \epsilon_t^2} \tag{3}$$

其中 MSE 用来检验回归数据集在整个数据集上的偏差程度和波动程度，DW 则用于检测线性回归模型中的残差是否存在自相关性。其中一次均值回归的 MSE 如图：发现部

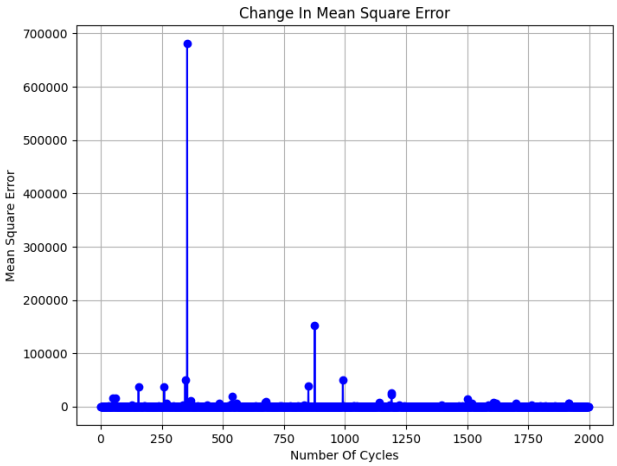
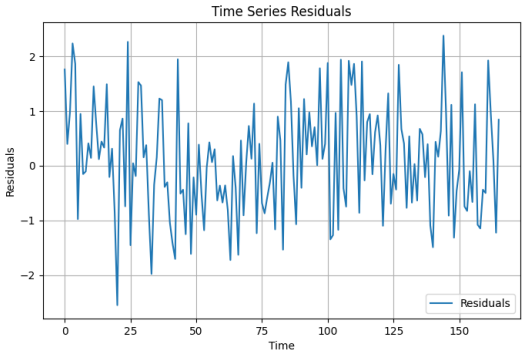
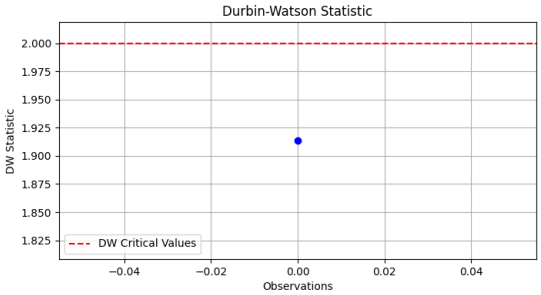


图 **4**　均方误差结果

分时间序列的 MSE 在较大程度上偏离数据集的 MSE 范围，我们考虑这种数据集在训练集上的基础数值过大，对其进行单独处理后重新进行回归。

对其中一个时间序列进行 DW 检验如下图：





基于 MSE 和 DW 的预测准确率检验反馈，我们不断调整和优化回归随机森林模型结构及相关超参数，同时过滤了数据集中的较大波动，获得了训练出针对电商平台需求量时间序列的回归随机森林模型。

5.1.6 回归预测

在经过波动过滤和超参数优化后，我们训练出针对电商平台需求量时间序列的回归森林模型，对 1996 个时间序列进行了均值回归，根据回归的结果进行各商家在各仓库的商品在 2023-05-16 至 2023-05-30 的需求量预测，并同时绘制了每个时间序列的回归与预测效果图。以 $seller_5 - product_{43} - warehouse_1$ 为例，其回归与预测效果图如图5。

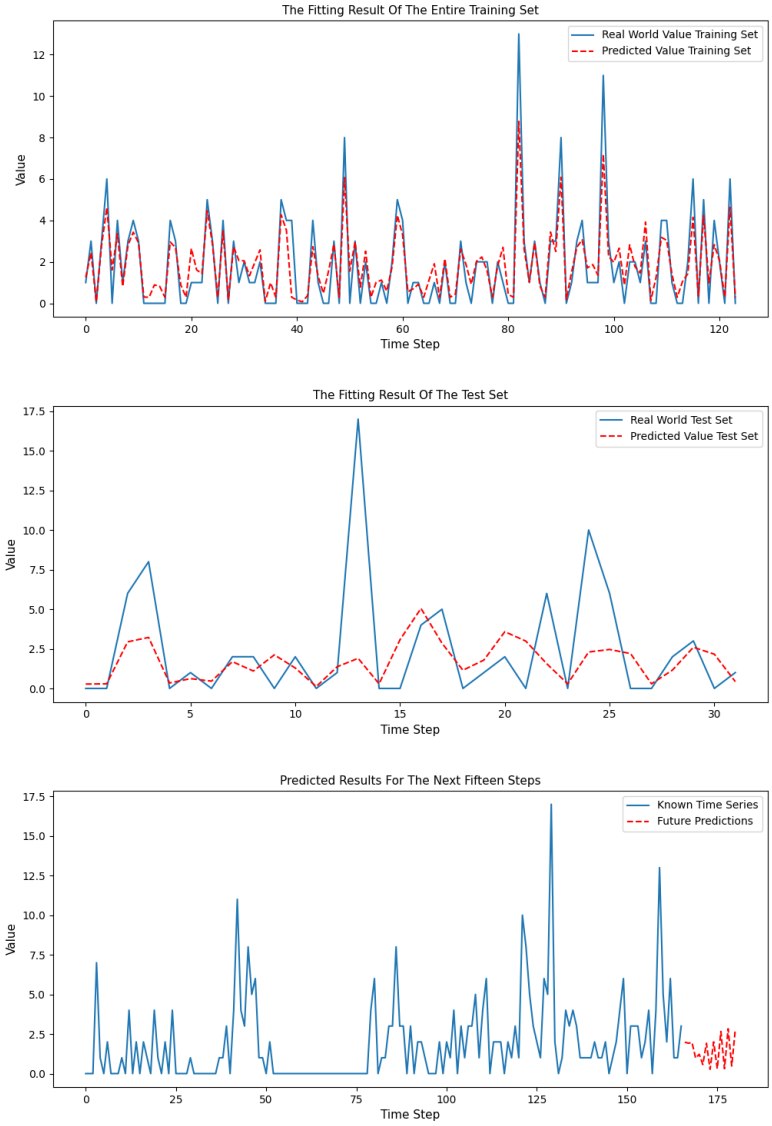


图 **5** 回归预测效果图，由上至下分别为拟合时间序列，测试集结果，预测效果，修改措辞

从图中可以看出，在整个训练集上，随机森林使用滑动窗口进行特征提取时，部分忽略了其波动较大的峰值，这一特征在验证集的回归图中体现更加明显：回归曲线在平稳范围内进行波动，并选择性地忽略了极大波动的峰值。在预测效果图中，我们可以看出 2023-05-16 至 2023-05-30 的需求量预测都在一个稳定的范围内进行波动，提供了一

个较为稳健的均值回归，反映了时间序列的波动回归点。

在问题一中，我们将此次回归预测的结果作为本题结果写入附件表格，因为我们认为，在较长时间跨度上进行回归分析，其得到的回归均值更有助于决策者对长期库存规划等决策进行更稳健地处理，以应对长期战略发展地需要。但也要注意到，我们在问题一后续的建模中会进行对时间序列波动规律的建模，不同商品的时间序列波动规律也应当是决策者进行决策时需要参考的维度。

5.1.7 历史波动随机变量

对于时间序列预测所需的波动情况也需要从已有的时间序列数据得到。我们认为，这种波动来源于市场的随机行为，可以用随机变量的统计分析进行描述。本问题中，定义逐差时间序列

$$D(T) = Q(T+1) - Q(T) - (Q(T_0 - 1) - Q(1)) \tag{4}$$

式中 T_0 为预测时间序列的第一天，$T_0 - 1$ 为现有时间序列的最后一天，减去最后一项使得整个时间序列均值为零，也即

$$\sum_{T=1}^{T_0-1} D(T) = 0 \tag{5}$$

对 $D(T)$ 作统计分析，可以得到逐差 D 的统计分布，预测中按照此统计分布抽样得到每天的逐差 $D_{pre}(T)$，叠加在随机森林预测得到的 $Q_{pre}^{rf}(T)$ 上作为波动修正，模拟出反映实际时间序列特征的变化趋势

$$Q_{pre}^{cr}(T) = Q_{pre}^{rf}(T) + D_{pre}(T) \tag{6}$$

作为库存规划的参考。计算中，我们限制以下条件：

$$Q(T_0 - 1) + \sum_{T=T_0}^{T_1} D_{pre}(T) \geq 0 \tag{7}$$

$$Q_{pre}^{rf}(T) + D_{pre}(T) \geq 0 \tag{8}$$

(7) 式中 T_1 为预测序列的最后一天。(7) 式代表如果时间序列按照此波动进行，不会出现预测值为负的情形，(8) 式则使波动修正的 $Q_{pre}^{cr}(T)$ 恒正。由 (5) 已知 $D(T)$ 之和为 0，再加上限制条件的影响，$D(T)$ 很好地代表了波动性而不包含多余信息。

5.1.8 时间序列的分类

聚类分析

聚类分析（Cluster Analysis）是一种无监督学习方法，旨在将数据集中的对象划分成不同的组（或簇），以使同一组内的对象彼此相似，而不同组之间的对象彼此不同。聚

类是一种探索性数据分析技术，用于发现数据中的隐藏结构，而不需要预先知道数据的标签或类别。

聚类分析中的一些关键概念和方法：

1. **簇**：簇是一组相似的数据点，它们在某种度量空间中彼此接近，但与其他簇的数据点不太接近。簇内的数据点应该具有高度相似性，而簇间的数据点应该具有较低的相似性。
2. **相似性度量**：聚类分析通常使用某种相似性或距离度量来衡量数据点之间的相似性或距离。常用的度量包括欧氏距离、曼哈顿距离、余弦相似度等。
3. **聚类算法**：聚类算法有许多不同的聚类算法，每个算法都有其独特的方法来确定簇。一些常见的聚类算法包括：

   K-Means：迭代聚类算法，通过最小化簇内数据点与其质心的平方距离来分配数据。

   层次聚类：根据数据点之间的相似性逐渐构建层次结构的簇。

   DBSCAN：一种基于密度的聚类算法，将数据点聚类成具有不同密度的簇。

   高斯混合模型（GMM）：一种使用概率分布来建模数据点分布的聚类方法。

   谱聚类：使用数据点之间的相似性矩阵来进行聚类。
4. **簇数目**：确定要分成多少个簇通常是聚类分析的一个关键问题。某些算法（如 K-Means）需要提前指定簇的数量，而其他算法（如 DBSCAN）可以自动确定。
5. **簇**：对聚类质量的评估通常需要使用一些指标，如轮廓系数、Davies-Bouldin 指数等。这些指标可以帮助确定哪种聚类解决方案更好。

基于 DTW 的改进 K 均值聚类分析

在现实情况中，企业会首先根据数据的历史情况，分析出需求量序列的数理特征，对相似的需求量序列进行归类，根据分类结果我们可以进行更加精确的预测。考虑到处理的数据集是时间序列这样一种特殊的数据集，我们采用动态时间规整（DTW）方法来分析时间序列之间的相似数理特征。DTW 方法可以在时间序列中找到相似的子序列，从而实现特征匹配；并且 DTW 对噪声和异常值的鲁棒性相对较强，可以较好的排除异常情况的影响。对 1996 个时间序列之间进行 DTW 距离计算，并将计算结果保存在矩阵中，得到大小为 1996×1996 的 DTW 距离矩阵 D_{dtw}：

$$D_{dtw} = \{d_{ij}\}$$

其中 d_{ij} 表示第 i 个时间序列和第 j 个时间序列的 DTW 距离。基于 DTW 距离矩阵来衡量时间序列之间的相似数理特征，我们对传统的 K 均值聚类分析进行优化和改进，使用 DTW 距离代替欧氏距离，并对一些模块进行了结构优化，使其在对时间序列数据集的聚类分析上拥有更好的针对性和精确性。从而得到了基于 DTW 距离的 K 均值聚类分析模型

$$D_{dtw} \to S_k : \{Q_1^k, Q_2^k, \cdots, Q_{n_k}^k\} \tag{9}$$

其中 S_k 为第 k 个聚类集合，n_k 为聚类中心的个数。

5.1.9 分类相似度检验

语义标签的数学化

生活经验中，同一类型、价格类似等具有各种相似属性的商品往往具有相似的出货量特征，其同时还受到售卖地区经济水平、库存分级等因素的影响。因此，在不考虑实际售卖数据，仅考虑诸如以上提到的时间序列语义标签，也能够对 $Q(T)$ 作出定性的分类。我们参照附件二三四，建立了语义标签的分类相似度检验模型。

基于语义标签的相似度

由“商家”、“商品”、“仓库”维度决定的每一时间序列 $Q(T)$ 都对应着“商品某级分类”、“商家分类”、“仓库区域”等八种语义标签，为 $A = [A_1, A_2, ..., A_8]$，相应的维度设为 $[X, Y, Z]$；每种标签下按序包含 2 种到 291 种不等的细分标签，数量为 a_{18}。对于任意的 A_i，我们使用 a_i 维的单位矢量描述，$Q(T)$ 对应的细分标签位置处取 1，其他位置取 0。例如，$Q_j(T)$ 维度 $[X, Y, Z]^j = [seller_{19}, product_{448}, wh_{30}]$，$\overrightarrow{A} = [$数码$, C, Large,$ 手机通讯,手机配件,手机配件$_{12}$, 中心仓, 华南$]$，数学化后有例如 $A_2 =$“C”，对应相应细分标签序列的第三个，因此 $\overrightarrow{A_2} = (0, 0, 1, 0)$。最终我们获得了数学化的时间序列语义标签描述数阵 $A = [\overrightarrow{A_1}, \overrightarrow{A_2}, ..., \overrightarrow{A_8}]$。自然而然地可以想到，当两时间序列 A_1,A_2 有 x 个标签相同时，定义 $F(A_1, A_2) = \sum_{i=1}^{8} \overrightarrow{A_1}\overrightarrow{A_2}$，显然有 $F(A_1, A_2) = x$。进一步地，当两个不同细分标签有一定相关性时，定义相关矩阵序列 $R_{sim} = \{R^1, R^2, ..., R^8\}$，$R^m = \{R^m_{ij}\}$，令 $F(A_1, A_2) = \sum_{m=1}^{8} \overrightarrow{A_1}^T R^m \overrightarrow{A_2}$ 任意矩阵元 R^m_{ij} 就描述了第 m 种语义标签序列中第 i 种与第 j 种细分标签的相关程度，正值越大表明越相关，负值绝对值越大表明越不相关、区别越大。语义标签分类的关键就在于确定每个相关矩阵的矩阵元。在本问题下，我们根据 $[X, Y, Z]$ 的实际情况结合经验确定了矩阵的生成方法。

矩阵元计算

附件 2 中，三级商品分类具有分支树状图的结构特点。矩阵结构上，三个矩阵具有如下特点：对低一级矩阵 R^{m_0} 作分块，分块矩阵能够一一映射到高一级矩阵的矩阵元 $R^{m_0-1}_{ij}$ 上。高一级矩阵 R^{m_0-1} 可以由 R^{m_0} 上相应的分块矩阵所有元素加上 $R^{m_0-1}_{ij}$ 完全代替。实际问题中，第三级的商品分类除编号外没有比第二级分类更详细的信息，因此 R^3 并不具有多于 R^2 的信息。综上所述，前三个相关矩阵中只需考虑 R^2 的生成问题。我们采用以下方法统一地生成其所有的矩阵元：考虑到特定维度下商品出货量时间序列特征与一系列实际因素有关，我们设置了平均价格 (a)，购买周期 (b)，受众人群 (c)，价格波动 (d)，存储周期 (e) 五种商品特征，根据经验设定每种细分标签对应的特征数值，并设定公式计算这些数值的相似程度，并进行加权求和来计算两种细分标签之间的矩阵元。其中平均价格和价格波动单位为元，购买周期与存储周期单位为年，受众人群数值表示比例。表2是我们参考电商平台数据人工设定的定性的特征数值实例：

category	平均价格	购买周期	受众人群	价格波动	存储周期
厨房卫浴	400	1	0.3	600	2
电工电料	20	4	0.3	10	2
洋酒	100	1	0.2	300	5
环境电器	300	2.5	0.2	300	1
进口食品	80	0.3	0.5	50	0.3
猫狗主粮	60	0.4	0.2	40	0.4
手机配件	40	0.4	0.95	30	1

表 **2**　商品分类特征数值表

以下是我们设定的特征相似程度计算公式，其基础是以相对差值大小判断是否相似:$1-\left|\frac{x_1-x_2}{x_1+x_2}\right|$，同时根据不同特征对商品出货量的影响进行修正，其中包含一些可供调整优化的参数：

$$s_1=\left(1-\left|\frac{a_1-a_2}{a_1+a_2}\right|\right)\sqrt{1-k_1\left(\frac{a_1a_2}{a_1+a_2}\right)^2},k_1=0.0001 \tag{10}$$

$$\tag{11}$$

强调均价越大，相关性越强

$$s_2=\left(1-\left|\frac{b_1-b_2}{b_1+b_2}\right|\right)\sqrt{1-k_2\left(\frac{1}{b_1+b_2}\right)^2},k_2=0.5 \tag{12}$$

强调购买周期越小，相关性越强

$$s_3=\left|1-\frac{f(c_1)-f(c_2)}{f(c_1)+f(c_2)}\right|,f(x)=\sqrt{(x-0.5)^3}+1,c_3\in(0,1) \tag{13}$$

强调受众人群很多或很少时，分别具有较高的相似性

$$s_4=\left(1-\left|\frac{d_1-d_2}{d_1+d_2}\right|\right)\sqrt{1-k_4\left(\frac{1}{d_1+d_2}\right)^2},k_4=5 \tag{14}$$

同 2，价格稳定的商品特征越相关。价格波动取与平均价格的相对波动比例

$$s_5=\left(1-\left|\frac{e_1-e_2}{e_1+e_2}\right|\right)\sqrt{1-k_5\left(\frac{1}{e_1+e_2}\right)^2},k_5=1 \tag{15}$$

同 2，存储周期短的商品销售特征更类似。获得以上商品特征类似程度后便可以加权获得相关矩阵的矩阵元

$$s_{ij}\sum_{k=1}^{5}m_k s_k, m_1=0.02, m_2=0.004, m_3=0.008, m_4=0.02, m_5=0.06 \tag{16}$$

式中加权值不同用以表明不同商品特征对时间序列特征影响程度的不同。

附件三中商家分类采用相同方法处理，库存分类、商家规模则人工设定。附件四中中心仓与区域仓在规模上有明显差别，矩阵中赋值较高。由于各地区经济发展水平的不同，仓库区域分为 {华东，华南}，{华中，华北}，{东北，西南，西北} 三类，类似附件二方法设置高一级矩阵的矩阵元。

利用上述算法，我们得到了聚类簇中任意两个时间序列的相似度检验模型 S-I，并根据检验结果指导基于 DTW 距离的 K 均值聚类分析中聚类中心个数 K 的再选择和优化，这个过程实现了一个反向传播 (BP) 优化结构：

$$S \Longleftrightarrow K$$

5.1.10 时间序列聚类簇的获得

在经过相似度检验和基于 DTW 距离的 K 均值聚类分析的超参数调整的 BP 优化过程之后，我们确定了聚类个数为 15，实现了对 1996 个时间序列的聚类分析，使同一聚类簇中的时间序列在需求上的特征最为相似，并获得了时间序列聚类簇 (Time-Series-Cluster)：

$$D_{dtw}, R_{sim} \rightarrow TSC^K : \{S_k : \{Q_1^k, Q_2^k, \cdots, Q_{n_k}^k\}\} \tag{17}$$

5.2 问题二的模型与求解

5.2.1 时间序列的归类

在实际生产过程中，常常会出现一些新的时间序列维度，如新上市的商品，或是改变了某些商品存放的仓库等。此时，往往我们对新的时间序列维度认识不够充分，新维度的时间序列长度也并不支持进行较精确的回归分析，在这种情况下，我们对新维度时间序列在问题一的建模基础上对附件 1 的时间序列进行归类，并借助归类后的聚类簇内的时间序列数据集为后续的回归分析提供数据上更充分的支持。

时间序列的归类过程我们依次考虑聚类模型和相似度检验模型，其基本过程如图：

考虑聚类模型

我们将每一个新维度的时间序列都视为一个新的聚类簇中心，对于每一个新维度时间序列，我们遍历附件 1 中的 1996 个时间序列，找到与其 DTW 最近的二十个时间序

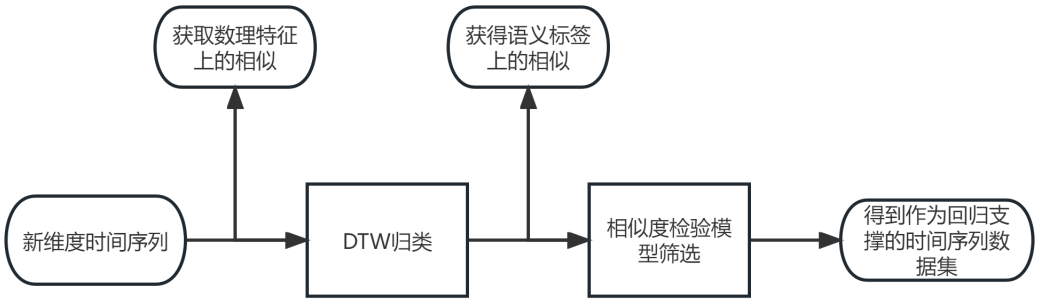


图 **6**　回时间序列归类图

列，以新维度时间序列为聚类簇中心进行归类。这些时间序列与新维度时间序列有着数理特征上的较大相似度。

以第一个新维度 $Seller_{19}, Product_{2215}, Warehouse_{21}$ 为例，其聚类簇内的 20 个时间序列与其的 DTW 距离如图7：可以看出选取的时间序列与聚类簇中心的 DTW 距离都

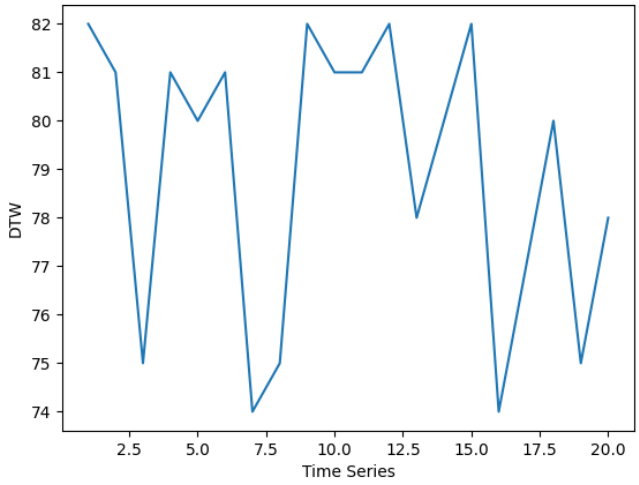


图 **7**　聚类时间序列数据集

在 74~82 的小范围内进行波动，说明基于 DTW 进行归类的数据集在数值上的表现较优。

考虑相似度检验模型

对于聚类中的二十个时间序列，我们考虑从语义标签的相似程度对其进行相似度检验，并从中选取语义标签相似度最高的十个时间序列。

$$R_{sim} : S_k \longrightarrow S_k^*$$

由此，我们基于每一个新维度的时间序列对附件 1 的 1996 个时间序列进行了归类，新维度的时间序列作为聚类簇中心，聚类簇中的时间序列与簇中心有着数理特征和语义特征两方面的相似，可以为簇中心在进行回归分析时给与数据上的参考和支撑。

5.2.2 回归时间序列数据集

考虑到附件五给出的时间序列作为回归分析的主要数据集，但其整体上时间较短，我们选择残缺时间序列补全和迁移学习的方法相结合。

其中残缺时间序列补全考虑在附件 1 的时间序列数据集上为附件 5 中的时间序列数据集进行扩展和补全，以更好的支持随机森林进行均值回归；迁移学习则考虑使用已在同类数据集上进行过大量训练和迭代调参的成熟回归模型直接对补全的时间序列进行均值回归，以提高回归精度和减少重复工程。其过程如图8所示：

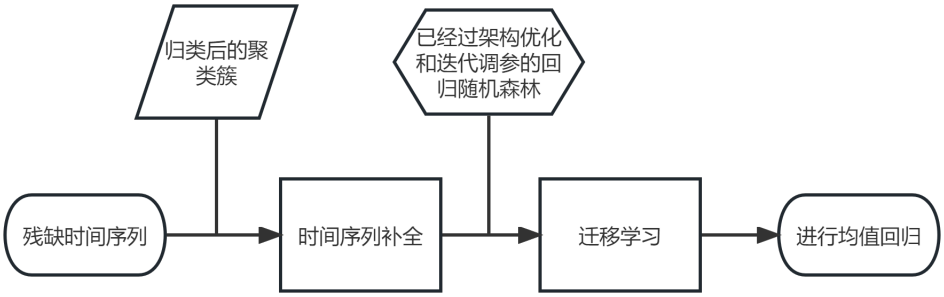


图 **8**　时间序列补全回归流程图

残缺时间序列补全

对于每一个新维度的时间序列归类得到的聚类簇，我们取聚类簇中的其它时间序列的平均值作为新维度时间序列在缺失部分的补全，补全得到的新时间序列既拥有了数据长度上的完整性以支持回归分析，又保留了原始数据作为回归分析的重要因素。以 9-2111-1 为例：其补全部分基本与原时间序列保持数理特征和语义标签上的一致。

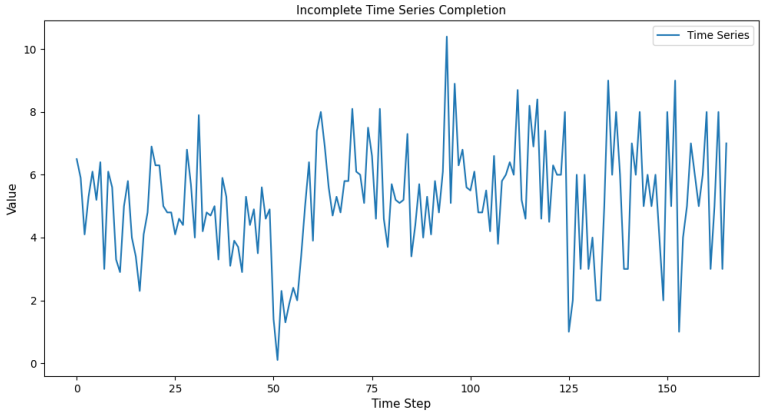


图 **9**　时间序列补全示意图

迁移学习

迁移学习（Transfer Learning）是一种机器学习方法，它涉及将一个已经训练好的模型（通常是深度神经网络）的知识或特征应用于解决与原始任务不同但相关的新任务。迁移学习的核心思想是，通过从一个任务中学到的知识，可以帮助改善在另一个相关任务上的性能，特别是在新任务的训练数据有限或昂贵的情况下。

考虑新时间序列在数理特征和语义特征两方面都与附件一中的部分时间序列相似，其回归分析的过程也得以借鉴。因此，我们采用迁移学习的方式，使用问题一中已经过架构优化和迭代调参的回归随机森林进行均值回归。

5.2.3 新维度的回归预测

我们根据回归随机森林的训练结果进行了新维度中各商家在各仓库的商品在 2023-05-16 至 2023-05-30 的需求量预测，并同时绘制了每个时间序列的回归与预测效果图。

以 $Seller_{11}, Product_{2160}, Warehouse_1$ 为例，其回归与预测效果图如图：

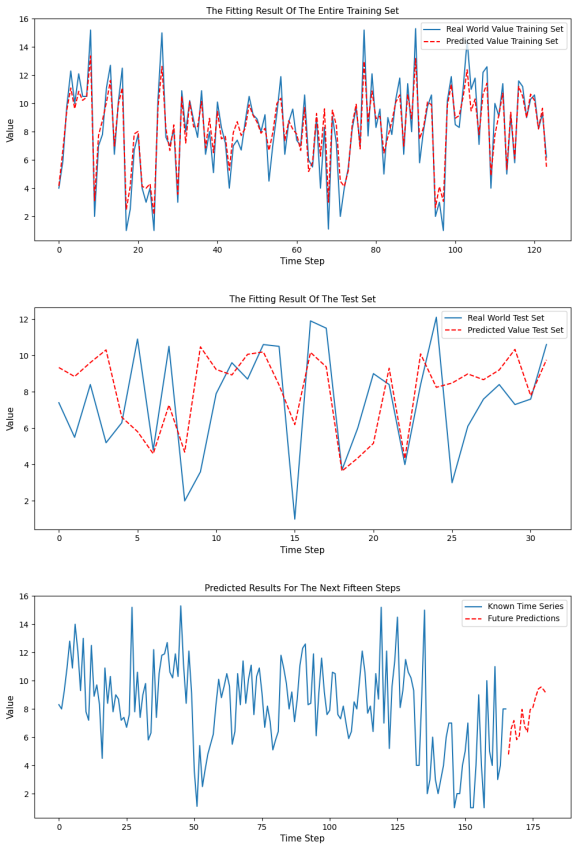


图 **10**　新维度回归预测示意图

可以看出经过补全的新维度时间序列经过回归随机森林的预测数据集在较小范围内进行波动，充分表现出其均值回归的特点。在问题二中我们将此次的回归结果写入结果附件表，我们认为，问题二的回归预测性质同问题一类似，其得到的回归均值更有助于决策者进行稳健的决策，并且新维度的时间序列发展趋势和波动周期尚未定型，因此回归均值更有助于在稳妥的决策基础上，依据于对新维度认识的加深而制定下一步决策，以应对多变的市场需求。

5.3 问题三模型的建立和求解

5.3.1 太阳黑子

太阳黑子是太阳周期性活动的一个基本标志，是太阳表面可以看到的最突出的现象。它存在于太阳光球表面，是磁场的聚集之处。其数量和位置每隔一段时间会发生周期性变化。太阳黑子历史变化周期如图11所示：”太阳黑子相对数”（Solar Sunspot Relative

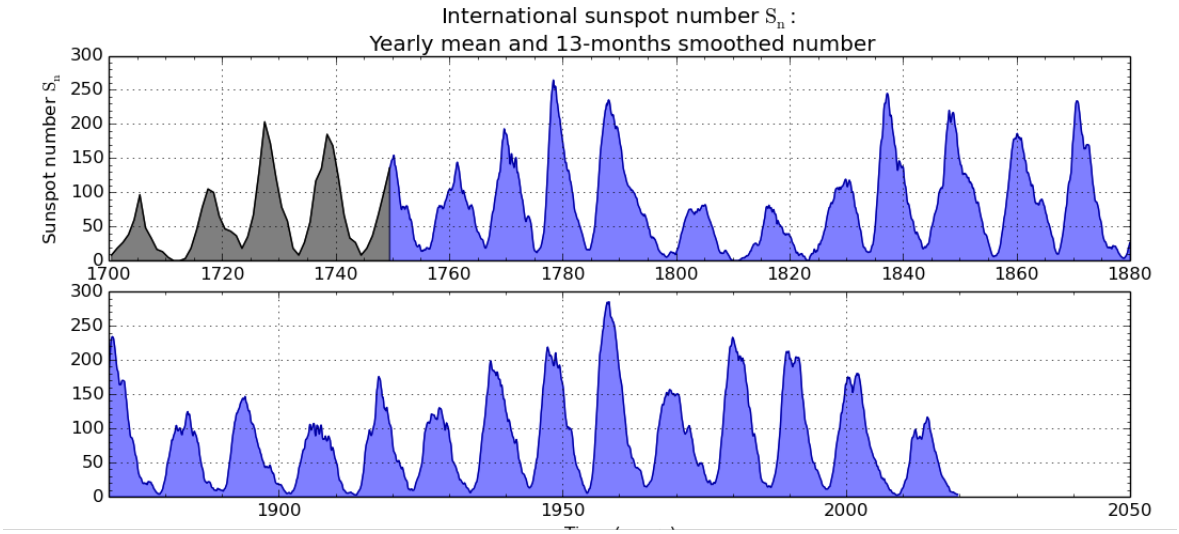


图 **11**　太阳黑子历史周期变化

Number）是用来描述太阳黑子活动的指标，通常是每月统计的。这个指标表示每个月观察到的太阳黑子数与历史平均值的比率。太阳黑子相对数通常用来监测太阳活动周期的变化。其公式如下：

$$Relative\ Sunspot\ Number(RSN) = K(10g + f)/(1 + G(10h + f))$$

其中：

1. K 和 G 是两个尺度因子
2. g 表示每月的太阳黑子组数
3. h 表示每月的太阳黑子的总数

4. f 是一个修正因子，通常取值为 1，这个因子用于考虑特殊情况和修正误差

太阳黑子相对数是一个表示太阳活动水平的指标，它将每月观察到的太阳黑子数量归一化到一个相对值，以便更好地理解太阳活动的变化。

5.3.2 电商出货量周期规律

我们考虑到出货量的时间序列波动趋势也类似于一种信号，在较大范围内拥有周期性变化的特征，其峰值的到达常常与外界因素有关，如“大型促销”等。我们考虑参考太阳黑子周期性变化规律中的研究方法，从去年双十一期间的时间序列数据集和近六月份的日常时间序列数据集提取电商波动相对指数 R，以更为精确地预测六月份的需求量峰值。一些时间序列的折线图如图：

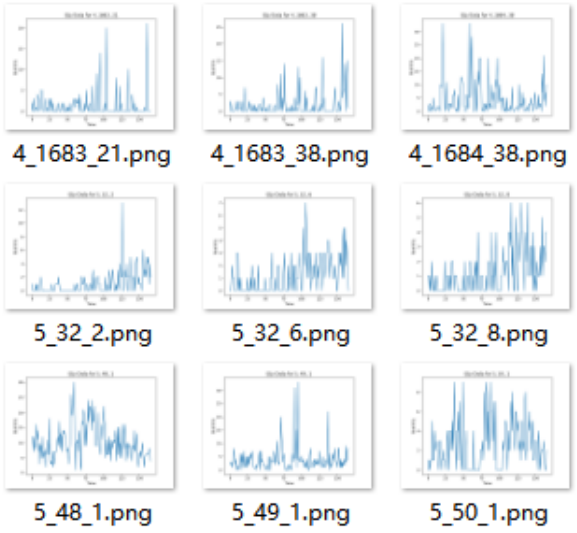


图 **12**　电商周期规律

电商波动相对指数

$$R = k(t \times g + f) \tag{18}$$

其中 g 为日常波动相对指数，f 为促销波动相对指数，t 为尺度因子，k 为修正因子，R 为所求的电商波动相对指数。

日常波动相对指数

日常波动相对指数通过采用历史趋势随机变量产生，统计分析逐差序列 $D(T)$ 相对于时间序列 $Q(T)$ 的相对标准差作为日常波动相对指数

$$g = H(Q^{(}T)) \tag{19}$$

其中 H 函数是根据逐差序列方法由时间序列获得描述其波动特征数值的函数。

促销波动相对指数

促销波动相对指数描述了双十一期间的波动和 6 月促销期间的波动相对大小。首先对双十一逐差序列统计分析获得其波动描述，然后分别按时间远近对日常数据加权求和，将和值作比，比值乘上双十一波动数据得到促销波动相对指数。

$$f = H(Q^{11}(T))\frac{\sum_i^N Q(i)q^{N-i}}{\sum_j^N Q(j)q^j} \tag{20}$$

其中 q 为衰减系数，这个值越大，促销波动对日常时间序列的依赖就越大。

现实情况调查

通过我们对现实相关情况的调查，我们发现许多电商平台在 6 月份伊始就已经开始了相关的活动促销，这是影响需求量的一个重要因素,也符合我们模型的假设。同时，我们发现绝大部分电商平台会在 6 月 18 日前后进行大规模的集体促销，故而绝大部分商品的需求量会在这一天前后达到一个峰值，所以我们考虑对叠加了电商波动相对指数的回归均值在 6 月 18 日前后加上一个强权，使其时间序列在 6 月 18 日前后的需求量数值有一个显著提升，从而提醒决策者对于这种情况进行充足的准备，在库存决策等方面为市场促销浪进行完善的考量和充分的处理。

5.3.3 叠加电商波动相对指数的均值回归

在获得电商波动指数和综合现实情况调查之后，我们建立了叠加电商波动相对指数的均值回归模型，并使用该模型对 2023-06-01 至 2023-06-20 时间段进行了回归预测。以 $Weller_{28}, Product_{808}, Warehouse_{41}$ 为例，其回归结果如图13：

可以看出，回归随机森林在验证集上过滤了波动趋势，而在最终的预测阶段，由于叠加了电商波动相对指数，其波动趋势在均值回归的基础上较大程度得反映出了波动趋势，并且在 618 附近给出了一个较大的峰值。在面对大型促销这种短暂的规律性波动规律，我们认为包含波动趋势的均值回归在决策过程中更有利于决策者为市场浪潮做出充分的准备，以一种更高的需求量预期进行决策，可以更好的应对突发情况和爆发性需求。因此我们使用叠加电商波动相对指数的均值回归模型对附件 6 其余时间序列维度进行回归，并将结果保存于结果附件表中。

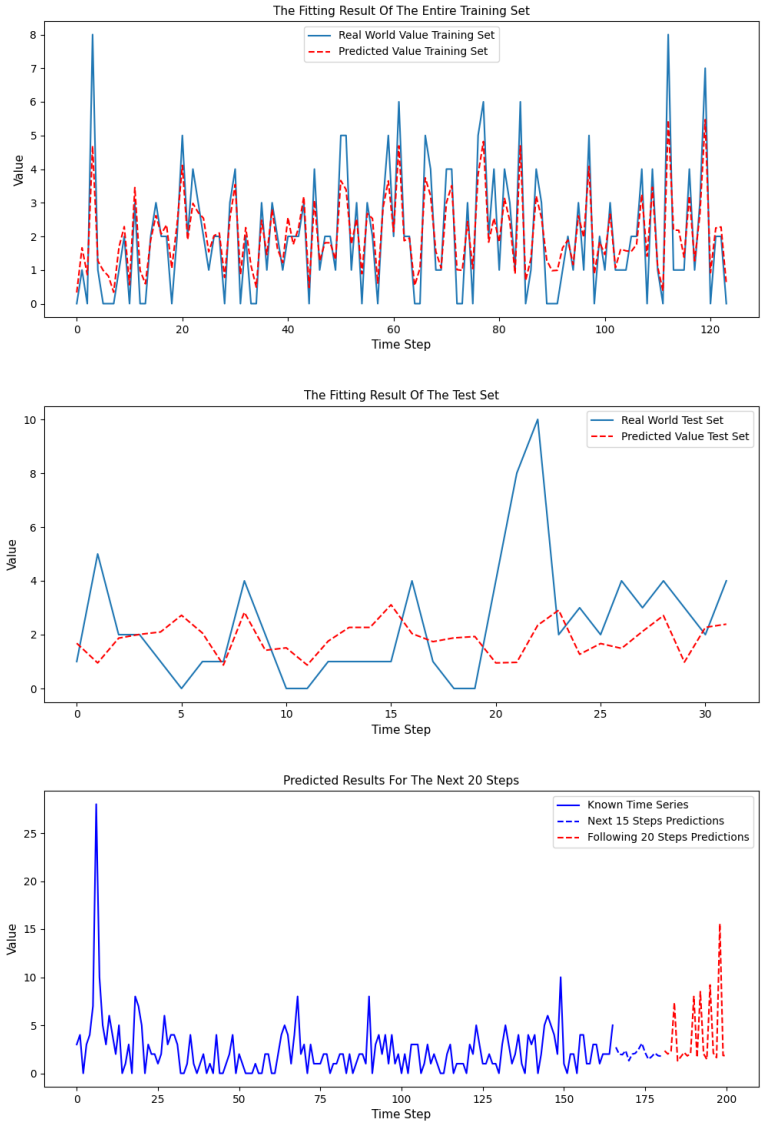


图 **13**　叠加电商波动的回归示意图

六、 模型的评价

6.1 模型的优点

(1) 充分地处理了大量复杂数据，对于每一个维度的时间序列都进行了充分的回归分析。

(2) 基于 DTW 的改进 K 均值聚类分析有效的处理了时间序列这种特殊的数据集的聚类问题。

(3) 考虑到了回归均值和波动规律两方面，为后续决策提供了多方面可供参考的数值。

(4) 面对不同场景采用了不同的建模侧重点，以应对复杂的市场变化。
(5) 创新点较多，引入了如 DTW、太阳黑子等研究方法来进行数学建模。

6.2 模型的缺点

(1) 对于数据的峰值部分的考虑不够充分，部分需要考虑的因素有所缺失。
(2) 处理较为复杂的波动时间序列的回归效果无法达到预期，在波动频繁的地方会产生较大偏差。
(3) 对于异常情况并没有建立起完善的应对机制，处理异常只能局限于部分场景。

6.3 模型的改进

(1) 对于历史趋势随机变量模型的建模加以完善，使其充分捕捉波动峰值。
(2) 引入优化算法如遗传算法等对模型的迭代和重构过程加以系统优化，使模型在数值表现上更加良好。
(3) 建立更加完善了异常应对机制，以处理市场变化中的各种异常情况。

参考文献

[1] 智能供应链: 预测算法理论与实战 [M]，北京：电子工业出版社，2023.

[2] Least Squares Quantization in PCM》，Stuart Lloyd，1957.

[3] Random Forests，Leo Breiman，2001

[4] Dynamic Programming Algorithm Optimization for Spoken Word Recognition，Sakoe and Chiba ， 1978.

[5] 刘润幸. 使用 SPSS 作多变量观察值的 ROC 曲线分析 [J]. 中国公共卫生, 2003, 19(9): 1151-1152.

附录 A 支撑材料内容

Analyse 文件夹：包含 Analyse Analyese4 代码是对原始表格数据进行数据提取、数据清洗、数据转化、数据分析的代码 Durbin-Watson 代码是进行 DW 分析的代码 Enhance_wmape 代码是进行 wmape 计算的代码

第一问至第三问主干部分代码：包含第一问至第三问的数据处理、格式转换、回归分析、特殊处理等全部完整过程代码

获取 DTW 距离代码：计算时间序列之间 DTW 距离的代码

基于 DTW 的改进 K 均值聚类分析代码

相似度检验模型代码

自定义第三方库代码：我们自定义的库，里面包含我们所有用到的自定义函数：

DTW_K 函数：基于 TDW 的计算分为 k 类的聚类分析函数

Enhance_wmape 函数：计算 wmape 的函数

DW 函数：DW 检验函数

DTM_k_6 函数：第二问所用的 k 均值聚类函数

Creat_Excel_1 函数：处理答案生成表格的代码

Similarity_test_function1 函数：相似度检验函数代码

get_data_dict 函数：获取附件 1 时间序列字典的函数

get_data_dict_5 函数：获取附件 5 时间序列字典的函数

get_data_dict_6 函数：获取附件 6 时间序列字典的函数

get_xyz 函数：获取时间维度列表的函数

get_xyz_tpl 函数：获取时间维度元组的函数

get_keys_by_value 函数：通过 value 获取 key 的函数

cluR 函数：计算电商波动相对数值 R 的函数