物流网络预测研究报告

目录

一、 选题背景与意义

1.1 背景资料

随着电子商务的兴起，物流快递业蓬勃发展，快递货量迅速增长。在物流网络中，分拣中心作为一个重要的中间环节，其管理效率将直接影响整个网络的履约效率和运作成本。如果能够有效地预测分拣中心未来一段时间需要处理的货量和进行合理的排班，将能够提升资源利用率并降低企业运营成本。因此，建立一个有效的物流网络预测模型具有重要的实际意义。

1.2 需要解决的问题

我们通过分析相关数据，运用数学思想，提出下列问题：

（1）根据已有历史数据，预测 57 个分拣中心未来 30 天及每小时的货量。

（2）根据分拣中心之间的关系，进行可视化分析。

1.3 选题意义

在当前竞争激烈的市场环境中，物流快递企业面临着提高服务质量和降低运营成本的双重压力。准确的货量预测不仅能帮助企业优化资源配置和人员排班，还能提升整体物流网络的响应速度和服务水平。此外，通过对未来货量和运输线路变化的预测分析，企业可以提前制定应对策略，减少突发事件对物流网络的影响，从而提高客户满意度和市场竞争力。因此，研究并构建高效的物流网络预测与优化模型，对企业的长期发展具有重要的战略意义。

二、 文献调研

物流网络预测涉及多种方法和模型，包括时间序列分析和机器学习方法。近年来，ARIMA 模型和 LSTM 模型在时间序列预测以及负荷预测中的应用较为广泛。ARIMA 模型擅长捕捉数据的线性趋势和周期性，而 LSTM 模型在处理复杂的非线性关系和长期依赖方面表现优异。Chun-Hua Chien 和 Amy J. C. Trappey (2021) 应用 ARIMA 和 LSTM 技术建立滚动预测模型，提高了需求和库存预测的准确性和效率，这些模型在实际应用中显示出比制造商的经验模型更优越的预测性能 [1]。

2.1 ARIMA 模型

ARIMA（差分整合移动平均自回归）模型是一种基于统计学的时间序列预测模型。它适用于线性、平稳的时间序列数据，能够通过差分和平滑处理将时间序列转化为平稳的序列，从而进行中短期预测。ARIMA 模型主要包括三个参数：自回归项（AR）、差

分项（I）和移动平均项（MA）。在物流领域，ARIMA 模型被用于预测物流量、货物运输需求等。龙宇等人（2023）在基于 ARIMA-LSTM-XGBoost 组合模型的铁路货运量预测中，利用 ARIMA 模型捕捉了铁路货运量的线性趋势和周期性变化。ARIMA 模型在物流预测中的应用主要具有以下特点：

1. 平稳性假设：ARIMA 模型假设时间序列数据是平稳的，或者通过差分处理可以变得平稳。这种平稳性假设使得模型在捕捉长期趋势时较为准确，但对短期波动的捕捉能力有限。

2. 线性特征：ARIMA 擅长处理具有线性趋势和周期性的时间序列数据；同样的，它的局限性在于面对非线性关系时，其预测效果较差。

在王代君等人（2024）基于 Bayes-ARIMA 的景区公路短时交通流量预测，利用了 ARIMA 模型进行景区交通流量预测，也指出了 ARIMA 模型在作为单一模型进行预测的缺点，即对非线性关系的捕捉能力差。ARIMA 模型中对历史数据的应用本质上也属于一种“先验信息”的运用，但这种运用是无条件、无去别的，这也就意味着所有数据都会参与到模型的拟合中，使得各种特殊情况被“折中”。[7] 并且 ARIMA 作为一种时间预测方法要求数据具有时间连续性，如果数据出现中断，那么该模型的预测性能将受到严重的影响。

针对 ARIMA 的优缺点，我们将引入下文的 LSTM 模型，对预测数据的准确性进行一个提升。

2.2 LSTM 模型

LSTM（长短期记忆网络）模型是基于 RNN（循环神经网络）的一种深度学习模型。相比于 RNN，LSTM 通过引入记忆细胞和门机制，有效解决了梯度消失问题，可以选择性地记住和忘记信息，从而更好地捕捉时间序列中的长距离依赖关系和非线性趋势。在物流领域，LSTM 模型被广泛应用于货物量预测、配送路径优化等。Hugo Tsugunobu Yoshida Yoshizaki 等 (2023) 在 54 个分销中心的案例研究中，LSTM 网络在 94% 的派送单位中表现优于统计方法，表明了 LSTM 在运输需求预测中的潜力 [3]：

1. 长距离依赖：LSTM 通过记忆细胞捕捉长距离依赖关系，能够处理具有复杂动态变化的时间序列数据。

2. 非线性特征：LSTM 擅长处理非线性关系，在面对复杂的时间序列数据时表现出色

吕志燕、王培进在基于 ARIMA-LSTM 的公路交通运输量预测（2023）[4] 一文中同样将 ARIMA 模型和 LSTM 模型进行结合，综合考虑了线性与非线性特征的公路运输量预测，并证明与单一的 ARIMA 模型和 LSTM 模型相比，ARIMA-LSTM 组合模型在公路交通运输量预测方面取得的效果更好。扬艳、黄晴等人发表的基于 ARIMA-LSTM 的货运量组合预测方法研究（2022）[6] 一文中指出 AIRMA 模型与 LSTM 模型相结合的

平均绝对百分比误差相较于单一模型降低了 51.45% 和 36.32%。有力证明了该模型的有效性。且近年来大量文献将 ARIMA 模型与 LSTM 模型相结合，在海运、空运等物流方面进行了预测，效果显著。结合以上研究成果，我们也将结合 ARIMA 模型及 LSTM 模型的特性完成上文预测需求。

三、 数据预处理

数据转换

对于附件一中的第一列数据，为避免造成二义性，同时便于进行数据挖掘，我们首先对这一列数据做如下操作，将“SC1”转换成“SC01”，“SC2”转换成“SC02”。同时将第二列中的日期类型转换为数值，以 8 月 1 日为基准作为第一天。

由附件 1，2 分别可知分拣中心过去 4 个月的每天货量，过去 30 天每小时的货量。首先对数据进行数据预处理。

缺失值的补全

通过对数据的筛选，发现附件 2 存在缺失值，结合问题背景，这是合理的，这代表此小时内分拣中心并没有开始工作，可能是没有货物运送往该分拣中心。对于缺失值，以数据 0 进行补充。

异常值处理

我们使用了 Matlab 软件进行正态分布检验。由于样本量大于 30，我们采用雅克贝拉检验 (Jarque-Bera test)。

对于一个随机变量 $\{X_i\}$，假设其偏度为 S，峰度为 K，那么我们可以构造 JB 统计量:

$$JB = \frac{n}{6}[S^2 + \frac{(K-3)^2}{4}]$$

可以证明，如果 $\{X_i\}$ 是正态分布，那么在大样本情况下 $JB - \chi^2(2)$（自由度为 2 的卡方分布）。如果结果发现 p 值大于 0.05 的水平，则接受零假设，这表明样本数据符合正态分布的要求。反之不符合，由于题目要对 57 个分拣中心进行货量预测，工程较大，同时为了全面展示整个预处理过程，这里我们仅给出部分分拣中心（SC1,SC36,SC57）的预测过程。

求解 P 值如下表 1 所示，SC1,SC36 未通过 P 值检验，SC57 通过，为便于直观理解，我们也绘制了 QQ 图。

表 1　SC1,SC36,SC57 p 值

分拣中心	SC1	SC36	SC57
P 值	1.0000e-03	1.0000e-03	0.1290

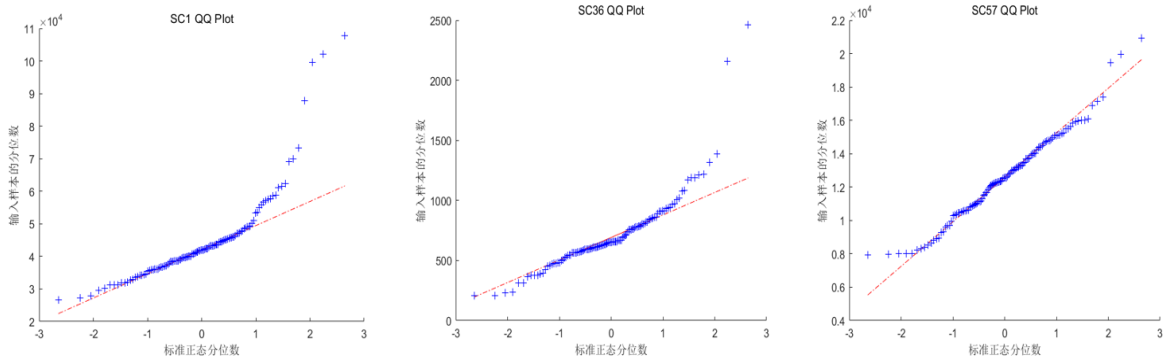

图 1　SC1,SC36,SC57 QQ 图

如果分拣中心的样本满足满足正态分布检验，我们则采用 3σ 检验，反之，我们采用箱线图对异常值进行检测。箱型图是一种用于显示一组数据分布的图表，通常定义 Q1 是第一四分位数,Q3 是第三四分位数,IQR 是四分位间距。其中“异常值”为小于 Q1-1.5IQR 或大于 Q3 +1.5IQR 的值。如下图 2 所示

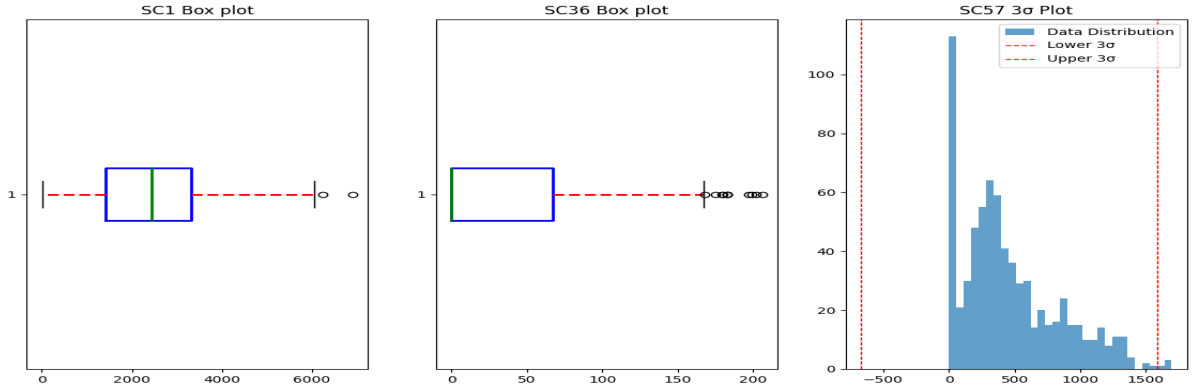

图 2　SC1,SC36,SC57 异常值检验图

根据结果，我们可以发现 SC1，SC36 不满足正态分布检验，SC57 满足正态分布检验，对异常值检验的结果我们发现均能够发现异常值。于是我们绘制了如下的时间序列图，其中 SC1,SC57 绘制在主坐标轴，SC36 绘制在次坐标轴，我们可以发现异常值的出现主要集中在 10,11 月，这主要是由于国庆假期，双 11 促销等节假日的影响。我们综合考量将双 11 等节假因素对自然月货量的影响视为合理因素，剔除了部分不合理的数据。

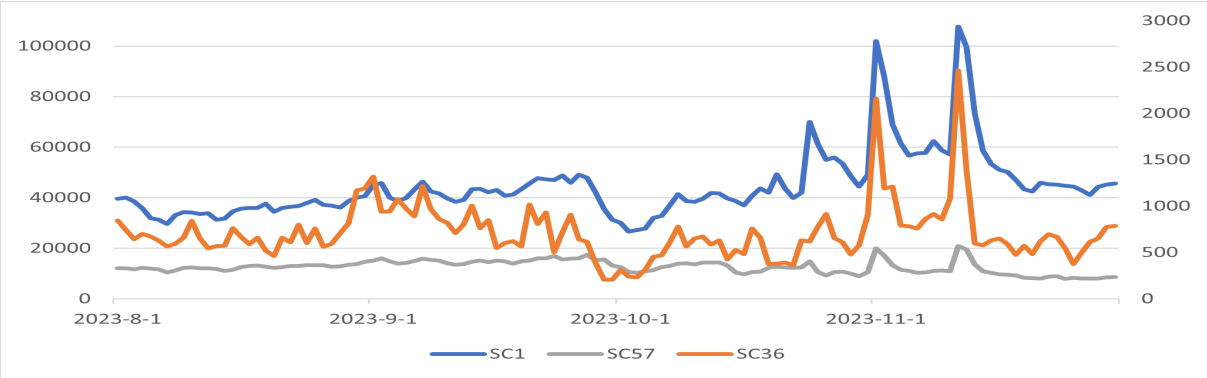

图 3　SC1,SC36,SC57 时序图

四、 第一部分：求解问题

4.1 线性回归

线性回归（Linear Regression）是一种基本的预测和分析方法，用于确定两种或多种变量之间的关系。在线性回归中，一个或多个自变量（也称为解释变量或特征）用于预测因变量（也称为响应变量或目标变量）。当这些变量之间的关系大致呈线性时，线性回归模型效果较好。以下为线性回归方程：

$$\begin{aligned} h(w) &= \sum_{i=1}^{n} w_i x_i \\ &= \theta^{\top} x \end{aligned}$$

假设现在真实的值为 y，预测的值为 h，损失函数公式为：

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - \theta^T x^{(i)} \right)^2$$

该公式计算了所有误差和的平方，也称为最小二乘法。损失函数越小说明误差越小，模型的泛化效果越好。

针对已有的数据，对其进行了一个线性回归模型拟合，分布结果如下：

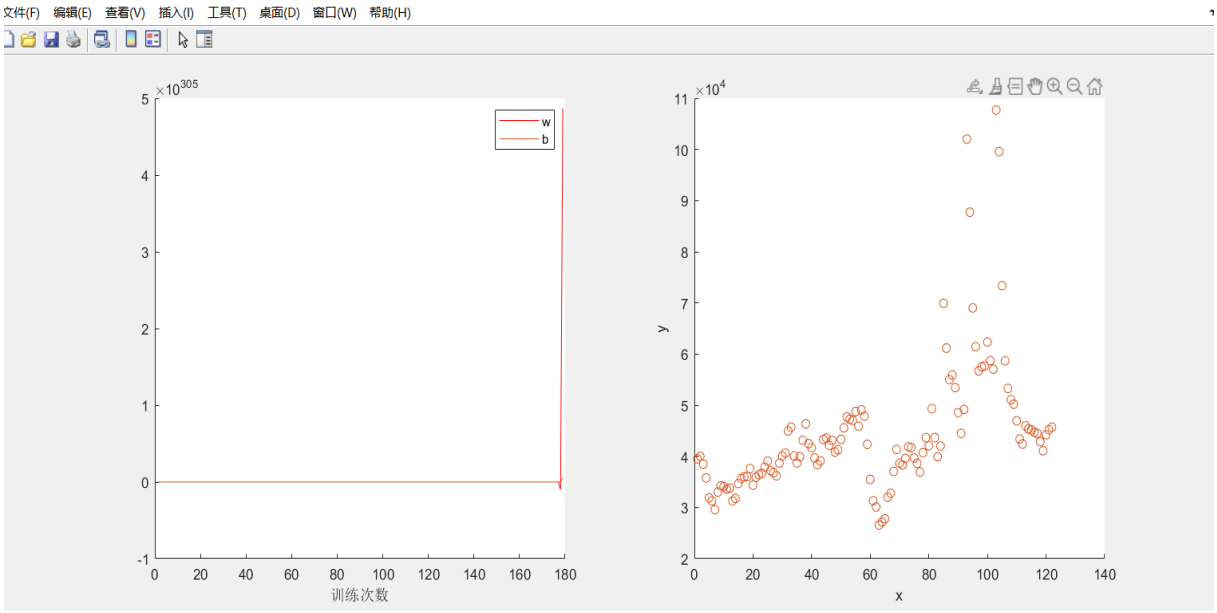

图 4　线性回归

经过对图示结果的仔细分析，可以发现当训练次数趋近于 180 时，线性模型 y=wx+b 中的系数 w 呈现出了显著的爆发式增长，而截距项 b 则保持相对稳定，未出现明显的变化。这种非线性的 w 值变化模式可能反映了数据在该区域的特定分布特性或潜在的非线性关系。在深入分析数据集的分布特征时，可以发现变量 x 在 [0,60] 的范围内呈现出一个明确的线性增长趋势。然而，当 x 的值超过 60 并接近 65 时，数据点却出现了显著的陡降，形成了明显的非线性拐点。此后，数据再次呈现爆发式的增长，这一变化模式显著地打破了单一线性关系的预期。因此，可以明确得出结论，该数据集的分布不符合线性分布情况。

4.2 bp 神经网络

bp 神经网络是一种多层前馈神经网络模型，其训练基于误差函数梯度下降的特征。该模型以信号的前向传播和误差的后向传播为特点，能够实现从输入到输出的任意非线性映射。BP 神经网络包括输入层、隐含层和输出层，形成三层网络结构。在正向传播中，输入信号经过隐含层的权重分配，传递至输出层进行输出值的计算。完成正向传播后，如果预测结果偏离期望误差，便利用反向传播调整权值和阈值，优化参数，从而建立最优模型。BP 神经网络因其自学习、自适应的特性适用于非线预测，在数据预测领域应用十分广泛。下神经网络的结构示意图

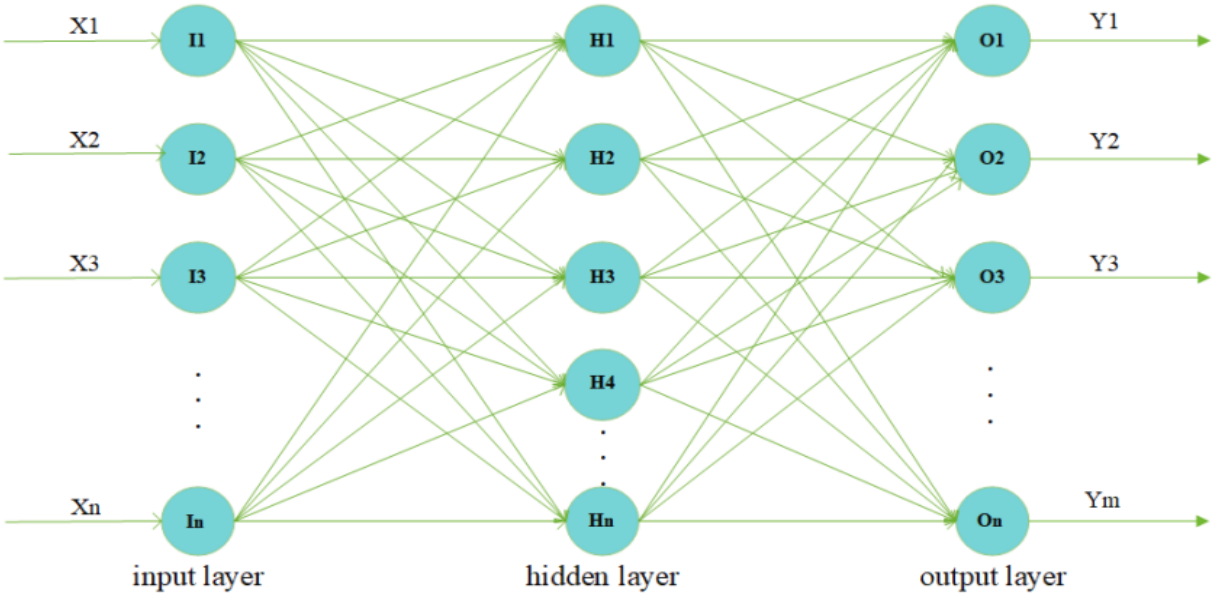

图 5　结构示意图

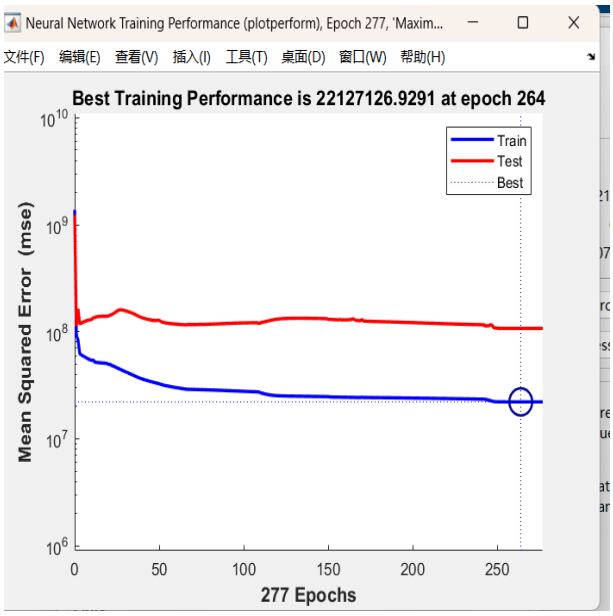

图 6　迭代图

本文模型的迭代次数为 277，由图可知，神经网络在迭代 264 次时收敛，误差过大，表现出预测模型反复训练中误差极大。由表可知，神经网络的拟合优度也不是很高，总体拟合优度约为 0.89，训练集和测试集的拟合优度都分别只有 0.93 和 0.75，这表明对数据的拟合效果较差

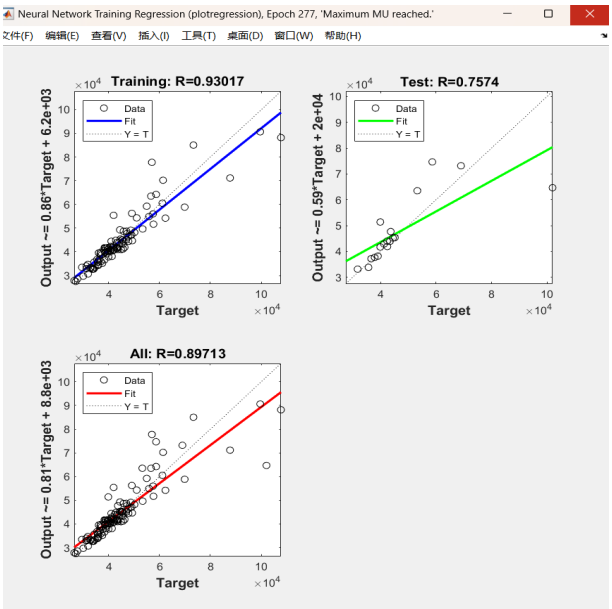

图 7　拟合结果图

单隐藏层的简单神经网络结构。

输入层有 13 个输入节点，对应于模型的 13 个输入特征。

隐藏层包含 10 个神经元。这些神经元可以捕捉输入特征之间的复杂关系。

输出层由单个神经元组成，适用于回归任务或二分类任务（根据具体应用场景而定）

图中显示了从输入层到隐藏层和从隐藏层到输出层的权重（标记为”W”）和偏置（标记为”b”），其中每个连接表示输入和神经元之间的加权关系。

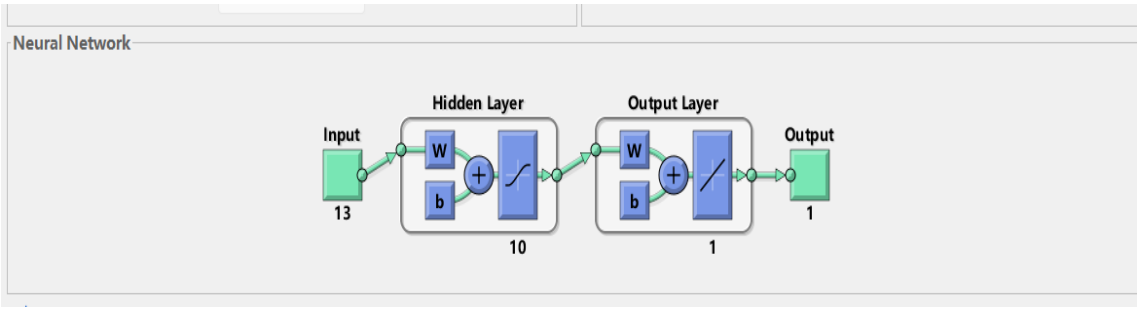

图 8　实际的结构图

下面是训练、验证和测试数据分配：

其中总样本数为 122。

训练数据占 70%，即 86 个样本。这些数据用于训练神经网络，通过调整网络权重以最小化误差。

验证数据占 15%，即 18 个样本。这些样本用于评估网络的泛化能力，并决定何时停止训练（以防过拟合）。

测试数据也占 15%，即 18 个样本。这部分数据在训练过程中不会用到，仅在训练完成后用来评估模型的最终性能，提供一个关于模型在未见数据上表现的独立评估。

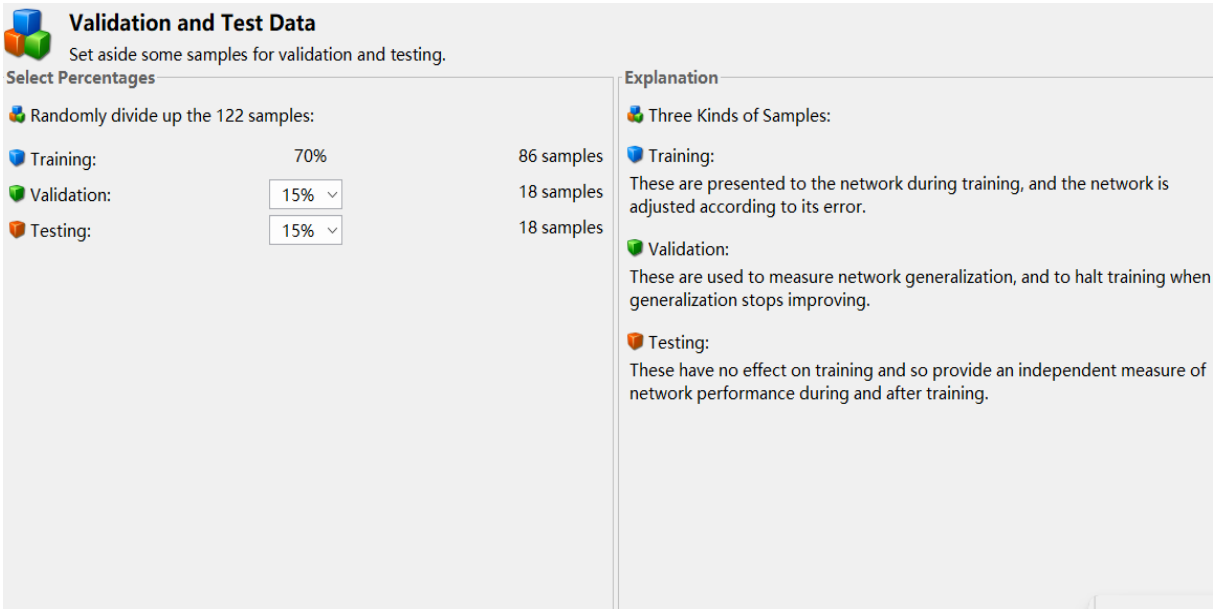

图 9　数据分配图

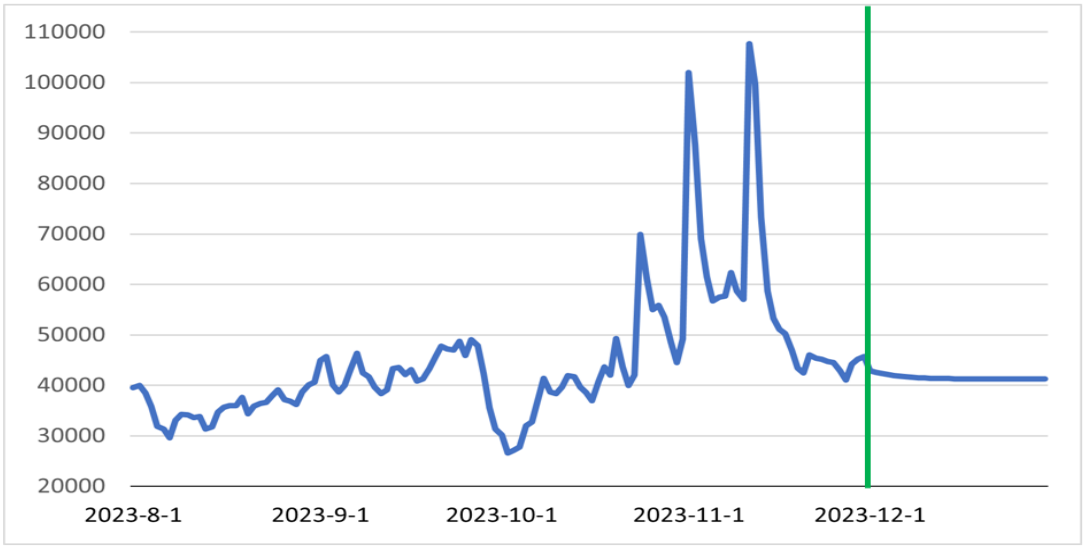

图 10　SC1 神经网络预测结果图

经过对拟合结果图的深入观察与分析，注意到在 2023 年 12 月 1 号之后，预测值呈现出一种线性递减的趋势。然而，根据现有的数据以及实际运营情况，分拣中心的货物量并不会呈现出稳定减少的态势。实际上，现实生活中每日的货物量会受到节日、销量、气候以及其他多种复杂因素的影响，从而产生大小不等的波动。这种平滑且单一的预估曲线在描述实际货物量变化时显得有效性较低。这种线性递减的预测模式通常基于一个

假设，即货物量以恒定的速率下降，这显然忽视了诸多可能影响货物量的外部因素及其变化性。因此，我们认为当前使用的 BP 模型（或其他类似模型）未能充分捕捉这些复杂的变化模式，从而无法对样本进行准确的拟合与预估。预测结果详见附件

4.3 ARIMA 时间序列预测

为对 2023 年 12 月分拣中心每天的货量进行预测，本问题基于 ARIMA 时间序列预测构建货量预测模型。

ARIMA 模型全称叫差分整合滑动平均自回归模型，该模型可以基于对时间序列数据的分析，包括分析趋势、季节性和随机性等特征，然后进行差分和平滑处理，将时间序列转化为平稳的序列，对未来中短期的时间序列进行预测 [5][8]。

步骤 1:White Noise 和 Stationarity 检验

White Noise 使一种频率恒定的随机信号，若序列是白噪声序列，则无法进行有效特征提取。

检验序列是否具有平稳性，若不具有，则需要进行差分。

步骤 2: 选择模型参数

ARIMA 模型由三个参数来描述时间序列，分别是 p、d、q:

参数 d 为差分的阶数，即需要几次差分。

p 是 AR 项，表示当前时刻的值与之前 p 个时刻的值之间的关系。

q 是 MA 项，表示当前时刻的值与之前 q 个时刻的噪声项之间的关系。

可以通过绘制 ACF 函数图像和 PACF 函数图像, 通过观察图像, 确定参数 p 和 q 可能的取值。

步骤 3: 模型检验

对拟合的 ARIMA 模型进行诊断，检查残差是否满足白噪声假设。使用残差得相关函数来检查模型的拟合情况，以验证模型的适用性。

步骤 3: 预测数据

使用得到的 ARIMA 模型进行预测 [8]

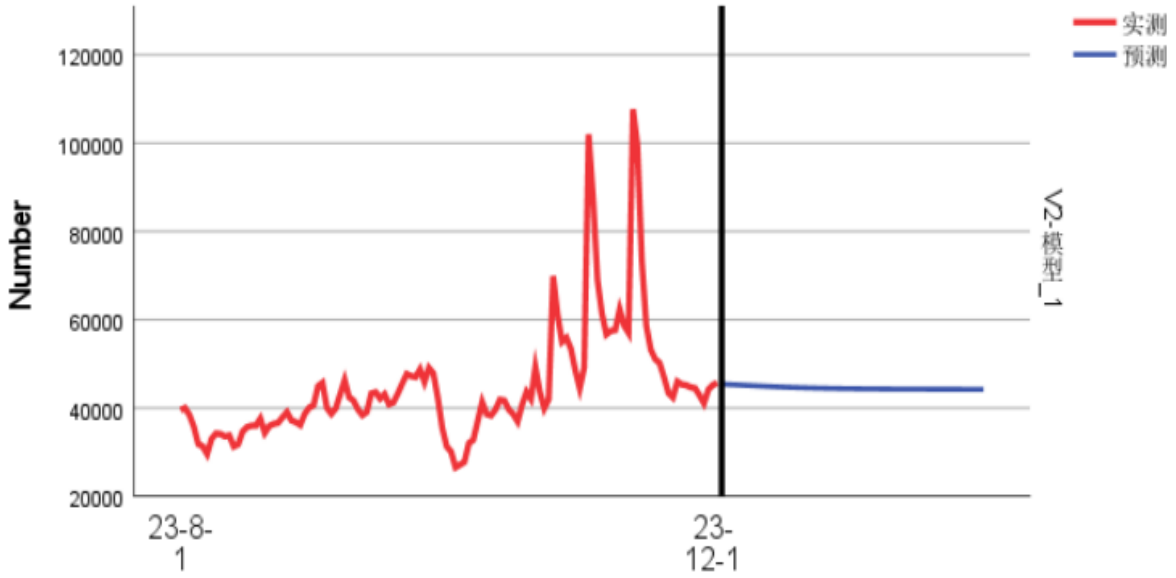

图 11　ARIMA 时间序列预测图

分析结果可以发现，ARIMA 预测未来 30 天的货量数据过于平滑，这不符合预测数据的预期。这是因为，ARIMA 模型的基本假设是时间序列数据是平稳的，或者可以通过差分等操作使其变得平稳。这种平稳性假设使得模型在预测时更多地关注数据的长期趋势，而可能忽略短期波动。因此，当时间序列中存在显著的短期波动或突发事件时，这体现在 10，11 月货量的显著波动变化，ARIMA 模型的预测结果可能会显得过于平滑，无法准确捕捉这些短期变化。其次 ARIMA 无法很好地捕捉非线性关系或复杂的时间序列动态。当数据中存在非线性趋势时，ARIMA 模型的预测结果可能会受到限制。为解决该问题，对于 ARIMA 的参数进行调整，仍然发现预测的结果拟合较差。

4.4 LSTM 长短时记忆预测

为弥补 ARIMA 的缺陷，我们引入 LSTM（长短期记忆网络），首先先简述该算法。

LSTM 是建立在 RNN 基础之上的，相比 RNN,LSTM 加入了一个记忆细胞，可以选择重要信息，过滤噪声信息，减轻记忆的负担。一个记忆细胞由三个门组成，分别是输入门 I_t, 遗忘门 F_t 和输出门 o_t，这三个门的值都在（0,1）范围内，它们的计算方法如下:

$$I_t = \sigma(X_t W_{xi} + H_{t-1} Whi + b_i)$$
$$F_t = \sigma(X_t W_{xf} + H_{t-1} Whf + b_f)$$
$$I_t = \sigma(X_t W_{xo} + H_{t-1} Who + b_o)$$

此外，还存在一个候选记忆元 $\tilde{\mathbf{C}} \in \mathbb{R}^{n \times h}$，它的计算方法与上述门类似，但是使用 tanh 函数作为激活函数，函数值范围为（-1,1），在时间步 t 处计算方法如下:

$$\tilde{\mathbf{C}} = tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c$$

输入门 I_t 可以控制采用多少来自 $overset{\sim}{C}$ 的新数据，而遗忘门 F_t 控制保留多少过去的记忆元 $C_{t-1} \in \mathbb{R}^{n \times h}$ 的内容，所以得出:

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{\mathbf{C}}$$

为了缓解梯度消失问题，并更好地捕获序列中的长距离依赖关系, 设置如果遗忘门始终为 1 且输入门始终为 0，则过去的记忆元 C_{t-1} 将随时间被保存并传递到当前时间步。

最后是隐状态 $H_t \in \mathbb{R}^{n \times h}$，由输出门控制，其值同样在（-1,1）内:

$$H_t = O_t \odot tanh(C_t)$$

只要输出门接近，就能够有效地将所有记忆信息传递给预测部分，而如果输出门接近 0，则只保留记忆元内的所有信息，而不需要更新隐状态。

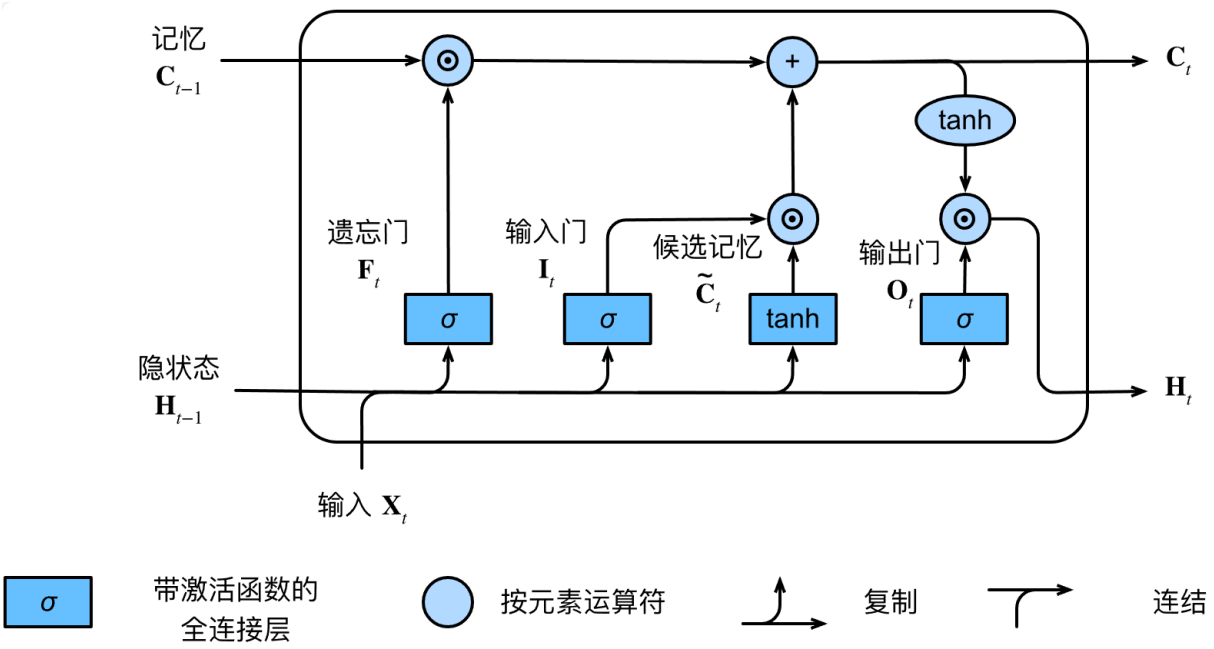

图 12　LSTM 模型流程图

4.4.1 模型调参

使用 matlab 中开源的 LSTM 预测工具包，需要调整的参数为学习率和隐藏单元数，以 0.0001,0.0005,0.001,0.002,0.005 为学习率集合，以 50,100,200,500,800,1000 为隐藏单元数集合，做笛卡尔积运算，寻找最优参数，结果如下

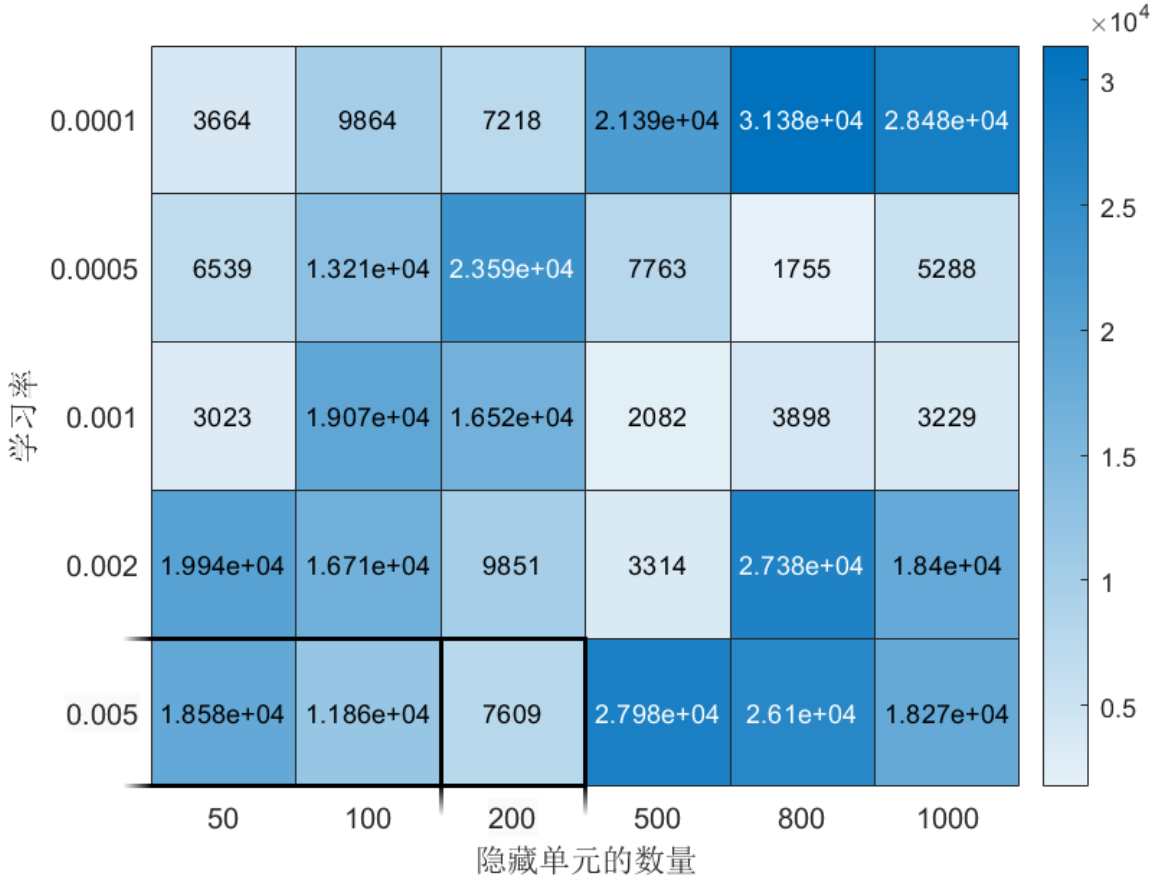

图 13　LSTM 调参 RMSE 图

当以 0.0005 为学习率，以 800 为隐藏单元数时模型的 RMSE 值为 1755，效果最好。

4.4.2 预测

由上一部分的结果，以 0.001 为学习率，以 800 为隐藏单元数量，设置丢失层概率 0.02，执行 100 轮开始预测。结果如下

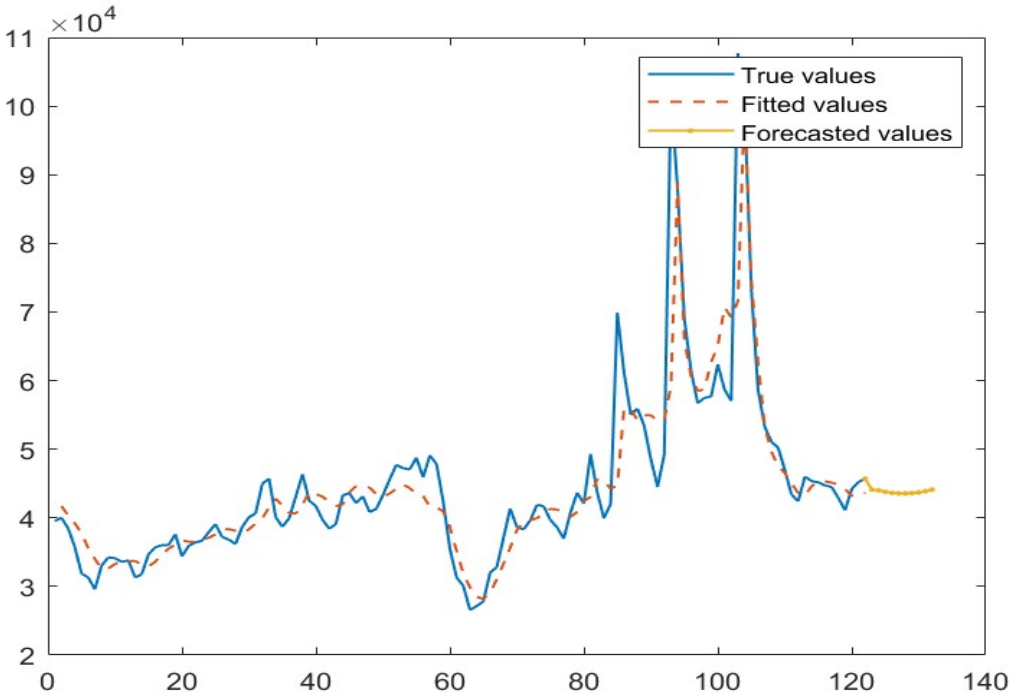

图 14　LSTM 模型流程图

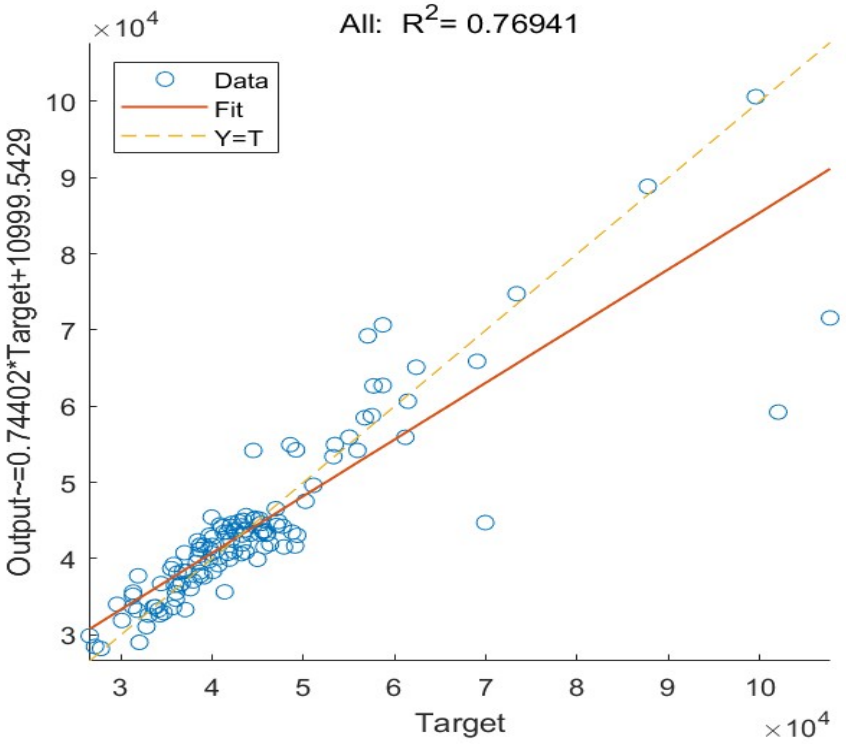

图 15　LSTM 模型 R^2 图

4.4.3 结果分析

我们选择了最优的学习率进行拟合，但所得的 0.76941 的 R^2 值并不算高，这表明模型只能解释 76.941% 的数据变异性，并没有达到理想的预测效果。但是需要更高的 R^2 值来确保模型的可靠性。当 target 在 3 到 5 的取值区间内，数据大量集中于拟合线附近，当 target 继续增大时，数据点基本上散布于拟合线两侧且距离较远，可见此时模型的拟合效果较差。考虑到这些拟合效果并不理想，我们决定在本研究中舍弃 LSTM 模型，并寻求其他可能更适合数据特性和预测需求的模型。

4.5 ARIMA-LSTM 组合模型

查找相关资料，我们发现 LSTM 能够很好的弥补 ARIMA 的缺陷，一方面它可以捕捉长期依赖关系，更好地捕捉时间序列数据中长期和短期模式。同时 LSTM 也是强大的非线性模型，能够更好的处理具有非线性关系的数据，结合 ARIMA 对于稳定和周期性模式的数据效果良好，采用 ARIMA-LSTM 组合模型。这帮助我们更好地应对复杂的数据特征。

在不同的物流分拣中心中，ARIMA 和 LSTM 在预测性能上有所不同。一般来说，ARIMA 模型具有稳定和周期性时间序列数据时效果良好，而 LSTM 在处理复杂的非线性关系和长期依赖时可能更具优势。，在不同的分拣中心中，选择适合的预测模型可以提供更准确的结果。

基于组合模型，需根据不同的分拣中心，对两个预测模型的权重进行分配，以使模型能够灵活的适应不同的数据预测，提高数据预测的准确性。

4.5.1 评测指标 RMSE

在模型评估与选择的过程中，选择恰当的评价指标至关重要。评价指标不仅能够帮助我们量化模型的性能，还能够指导进行模型的优化和选择。在销量、需求预测邻域，我们常用 RMSE, MAE 等作为模型的评价指标，而 RMSE（Root Mean Square Error，均方根误差）因其独特的优势而被广泛应用于各种预测模型中。

RMSE，全称为均方根误差，是一种衡量预测值与真实值之间差异的常用指标。其计算原理为：首先计算预测值与真实值之差的平方，然后对所有差值平方进行平均，最后取平均值的平方根。数学公式可表示为：

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(predicted_i - actual_i)^2}$$

其中，$predicted_i$ 表示第 i 个预测值，$actual_i$ 表示第 i 个真实值，n 表示样本数量。很多

算法使用 MAE 为标准进行优化，此处也给出 MAE 评价指标

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

其中，y_i 表示真实值，$\hat{y}_i$ 表示预测值，n 是样本数量。对比二者发现，RMSE 给每个错误的权重是不同的，error 越大的数据给的权重越大，这说明只要有一个非常不好的预测结果，整个 RMSE 都会很差。并且 MAE 优化的是中位数，而 RMSE 优化的平均值。结合本题背景，需要模型体现实际生活中的波动变化，这说明模型要对异常值具有一定的敏感性，避免出现上文 bp 模型、ARIMA 模型中平滑的预测结果，因此此处选择 RMSE 作为接下来的拟合预测模型的评测指标。

4.5.2 未来 30 天每天货物量预测

计算权重

RMSE（均方根误差）是衡量预测值与实际值之间差异的一种常见指标。其计算公式如下：给定预测值 $\hat{y}_i$ 和对应的实际值 y_i，共有 n 个样本，则 RMSE 可通过以下公式计算：

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

RMSE 的计算过程包括以下步骤:

首先计算每个样本的预测值 $\hat{y}_i$ 与对应的实际值 y_i 之间的差值，其次将每个差值取平方，接着求所有平方差的平均值，最后对平均平方差取平方根，得到 RMSE。

RMSE 越小表示模型的预测效果越好，因为它衡量了预测值与实际值之间的平均偏差的大小可以用 RMSE 来衡量组合模型的权重。

$$W_{ARIMA} = \frac{1}{RMSE_{ARIMA}} \Big/ \left(\frac{1}{RMSE_{ARIMA}} + \frac{1}{RMSE_{LSTM}}\right)$$
$$W_{LSTM} = 1 - W_{ARIMA}$$

其中，W_{ARIMA} 和 W_{LSTM} 分别表示 ARIMA 模型和 LSTM 模型的权重，$RMSE_{ARIMA}$ 和 $RMSE_{LSTM}$ 分别表示 ARIMA 模型和 LSTM 模型的 RMSE 值。对各个物流分拣点，两个算法的 RMSE（均方根误差）如下图 6，7 所示：

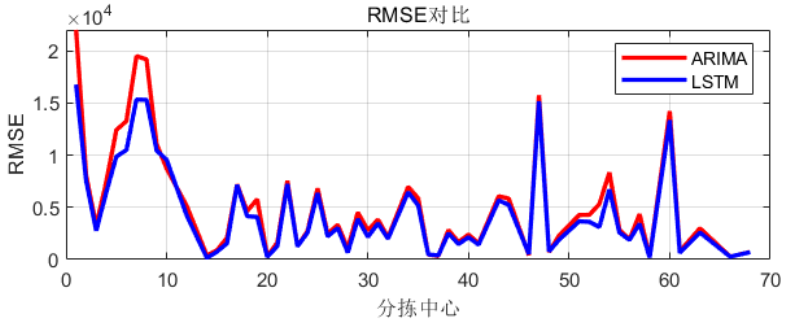

图 16　RMSE 对比图

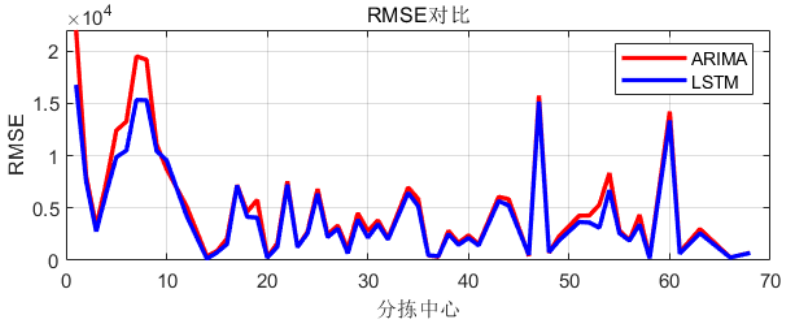

图 17　权重对比图

计算得各分拣节点的权重分配如下表 2 所示：

表 2　权重分配部分图

分拣中心	ARIMA	LSTM	分拣中心	ARIMA	LSTM
SC1	0.6689	0.3310	SC2	0.5460	0.4539
…	…	…	…	…	…
SC66	0.6339	0.3660	SC68	0.6128	0.3871

依据权重分配，对未来 30 天货量进行预测，如下图 8 所示

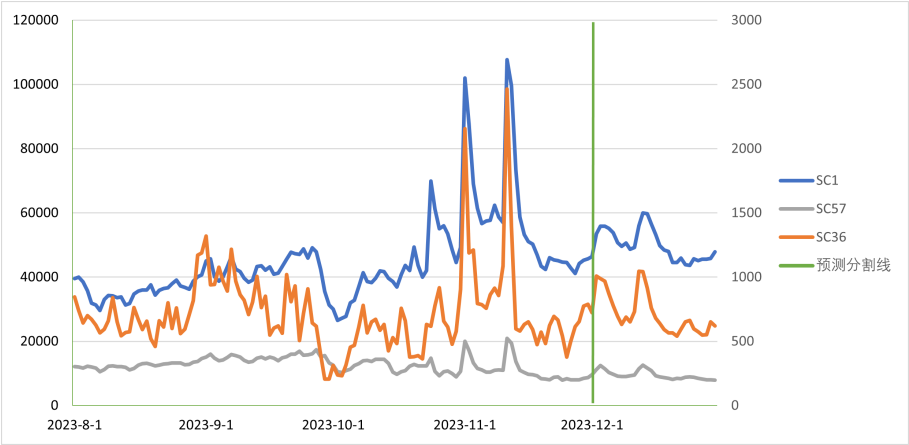

图 18　未来 30 天每天货物量预测图

分析图可以看出，对于未来 30 天预测的货物量前后的分布一致，满足了 3σ 检验或者箱线图检验，预测结果较为合理。同时，12 月初货物量的显著波动也符合 12 月初大购物的普遍规律，这能够说明对未来的货物量能够进行良好的预测。

通过结合 ARIMA 和 LSTM 模型，重复利用了两者在不同情况下的优势，提高了整体预测的精度和模型泛化能力。相比于上文 ARIMA 单模型的应用，LSTM 模型在捕捉时间序列中的长短期依赖关系方面表现出色，能够有效处理数据中的短期波动，避免了平滑的预测结果，提高了预测的可靠性。同时，ARIMA 模型在处理时间序列数据时，会进行差分和平稳化处理，可以有效消除数据中的非平稳性，并且较好地捕捉数据中显著的线性趋势和周期性，将其引入到后续的 LSTM 模型中，从而提高 LSTM 模型的训练效果和预测精度。通过 ARIMA 预先处理和预测部分数据特征，LSTM 模型可以专注于处理剩余的复杂非线性关系，减轻了 LSTM 模型的计算负担，在提高预测准确性和稳定性的基础上，提高整体模型的效率。

4.5.3 未来 30 天每小时货量预测

采取同样的方法对未来 30 每天每小时货量预测，首先得到两个模型的预测权重，结果展示部分如下表 3 所示。

表 3　权重分配部分图

分拣中心	ARIMA	LSTM	分拣中心	ARIMA	LSTM
SC1	0.4754	0.5245	SC2	0.4963	0.5036
SC3	0.5464	0.4535	SC4	0.5169	0.4830
…	…	…	…	…	…

下面给出部分预测结果图, 将未来 30 天每天及每小时的货物预测数据合并为一个表格，如下表所示。具体的内容详见结果 1 和结果 2。为便于直观展示每天每小时货物量的预测结果，这里仅仅给出 SC1 货物量预测图。

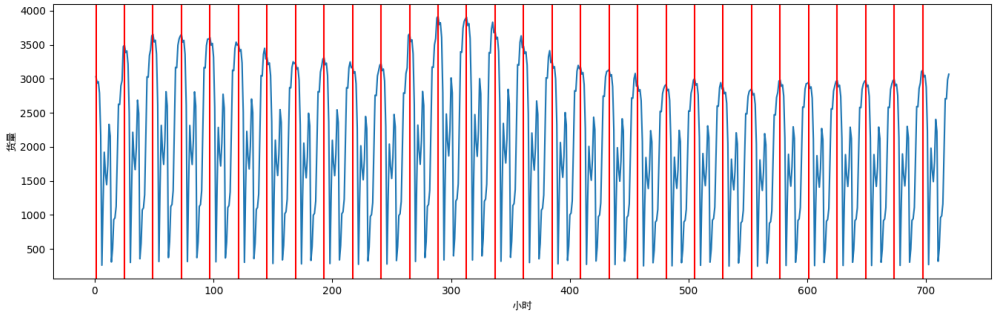

图 19　SC1 每天每小时货物量预测

4.5.4 预测结果

表 4　部分预测

分拣中心	日期	货量	分拣中心	日期	小时	货量
SC1	23-12-1	53448	SC1	23-12-1	1	3013
SC1	23-12-2	55901	SC1	23-12-1	2	3031
SC1	23-12-3	55866	SC1	23-12-1	3	2937
...	...	...	...	...	...	...

在每天和每小时的预测数据中，能清晰看到高峰期和低谷期的分布。整体货物量的趋势较为平稳，而在月末和月中可以明显观察到货物量较平常有明显提升，预测结果显示出一定的波动性，这与实际物流活动中的不确定性相吻合，证明了模型在处理短期波动数据方面表现出色。

五、 第二部分：可视化分析

根据附件内容以及第一部分做出的结果，发现不同分拣中心，不同日期，不同小时内货量存在差异。从现实社会中，对于货物的输送，理论上也应该存在各自运输路线的连通关系，基于此我们开始可视化分析分析。

5.1 连通图

根据附件 3 的内容，使用 Gephi 绘图工具做出分拣中心之间的子图

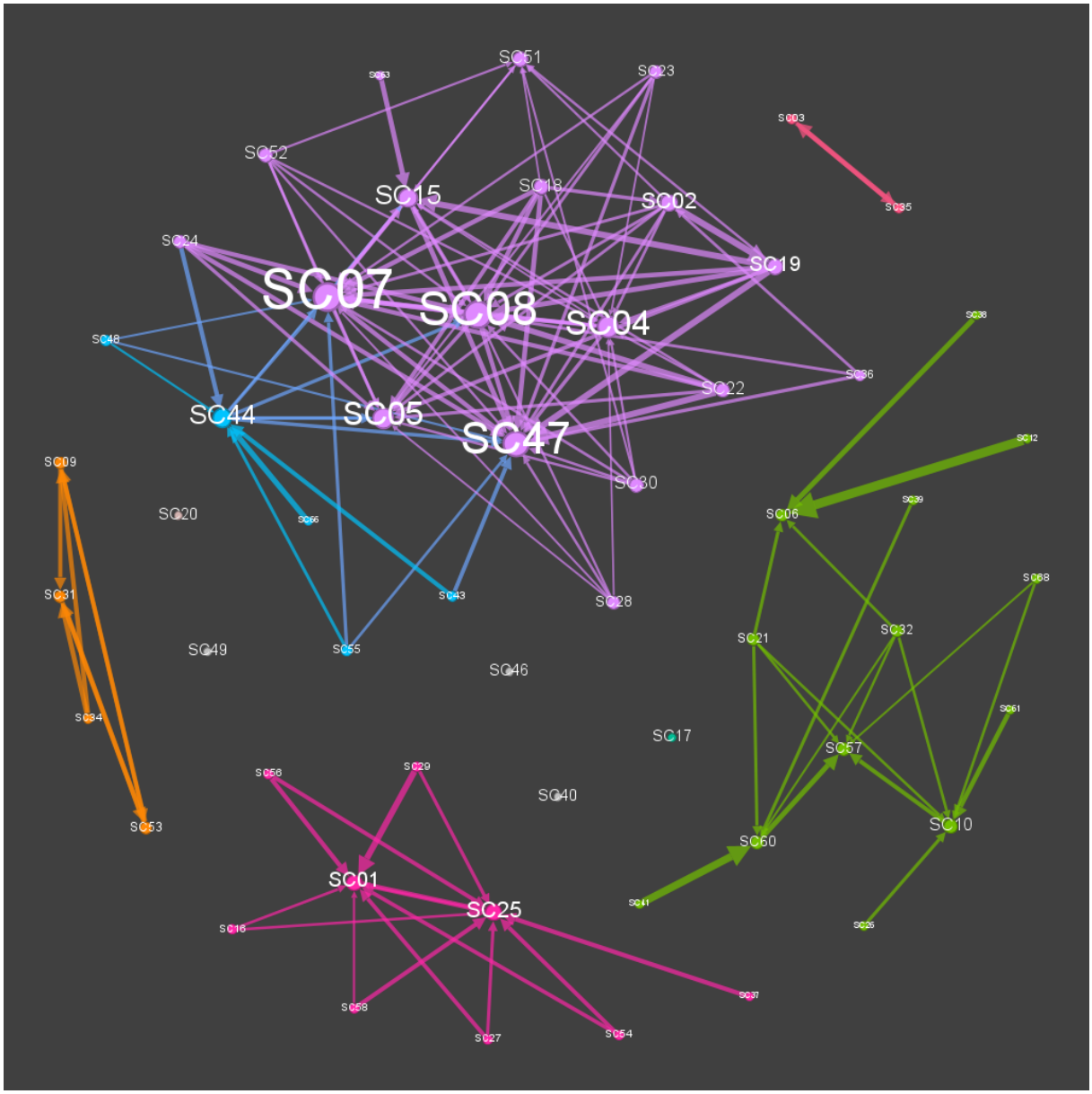

图 20　连通图

5.2 k-means 聚类

K-Means 算法是一种无监督学习，同时也是基于划分的聚类算法 [2]，一般用欧式举例作为衡量数据对象间相似度的指标，相似度与数据对象间的距离成反比，相似度越大，距离越小。相较于其他的聚类算法，K-Means 算法以效果较好、思想简单的优点在聚类算法中得到了广泛应用。算法需要预先指定初始聚类数目 k 以及 k 个初始聚类中心，根据数据对象与聚类中心之间的相似度，不断更新聚类中心的位置，不断降低类簇

的误差平方和（Sumof SquaredError，SSE），当 SSE 不再变化或目标函数收敛时，聚类结束，得到最终结果。

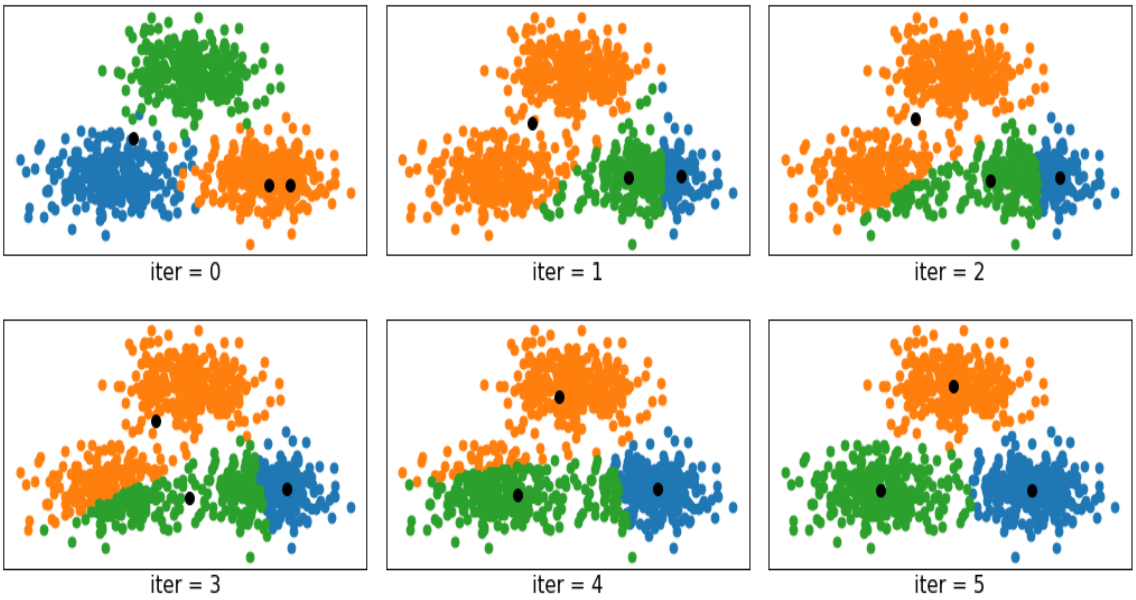

图 21　K-Means 算法迭代过程

我们通过 spss 求解得到聚类数目以及谱系图，详见附件，最终得到聚类结果如图 22 所示。我们设定 k 为 3 类，下图横坐标为分拣中心，纵坐标为分拣中心对应的货物量，聚类的颜色代表分类类别。

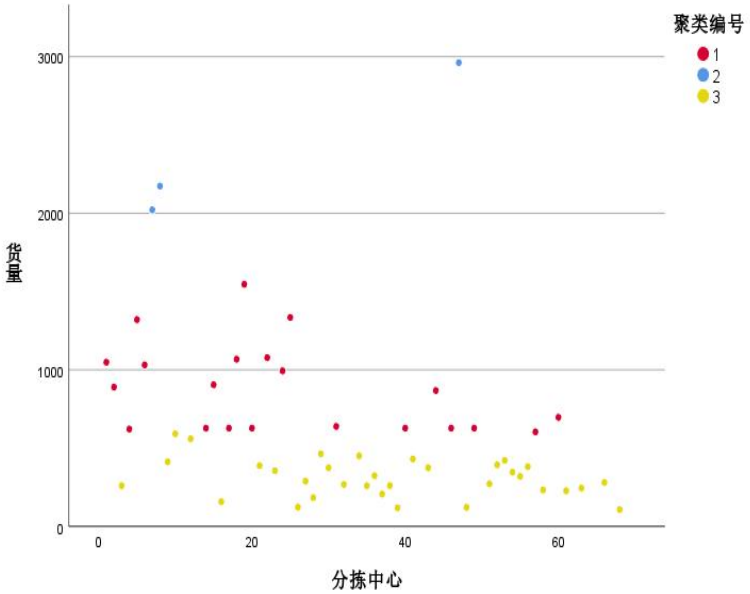

图 22　平均货量聚类图

5.3 系统聚类

有标准化和欧式距离

系统聚类是一种基于相似性或距离的聚类方法，它通过计算两类数据点间的距离，对最为接近的两类数据点进行组合，并反复迭代这一过程，将数据点逐渐合并成越来越大的簇，形成一个层次结构。

假设数据共可以分成 i 类：G1，G2...Gi, 用 Pi 表示 G1 中的第 i 个数据，Qi 表示 G2 中的第 i 个数据，则欧几里得距离的计算公式如图所示：

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\ldots+(p_n-q_n)^2}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}$$

我们通过 spss 求解得到聚类数目以及谱系图详见附件，最终得到聚类结果如图 23 所示。为对比两种分类的区别，同样设置为 3 类分类。

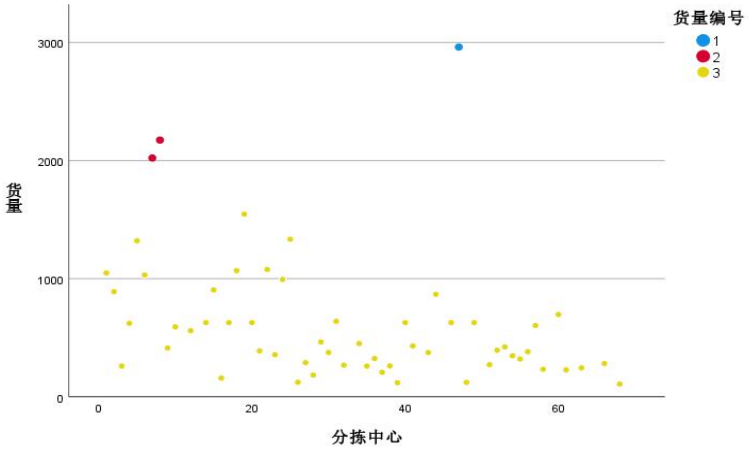

图 23　平均货量聚类图

5.4 货物运输连通子图

结合附件 3 的运输路线，我们绘制了连通图如下图 24 所示。分析图标我们可以发现，分拣中心的货物输送呈现出复杂的网络结构，并非是简单的线性或者树状结构，其中一些分拣中心 SC2，SC33，SC16 等作为关键枢纽，影响后续分拣节点的运送情况，决定整个货物输送的稳定性和连通性。最后我们还可以发现，输送路径呈现出多样性，输送路线更加灵活和鲁棒，这使得后续改变输送路线能够继续维持其稳定性。

图 24　连通子图

六、 总结展望

6.1 模型的优点

1.ARIMA 模型在捕捉线性趋势和季节性方面表现优秀，而 LSTM 在处理复杂、非线性的时间序列数据上具有较强的能力。ARIMA-LSTM 混合模型能够结合两者的优点，同时捕捉数据中的线性和非线性特征，提高预测准确度。

2. 考虑了周末节假日和特殊情况对于货量的影响。ARIMA 模型擅长捕捉时间序列中的周期性变化，LSTM 模型可以记住时间序列中的长期和短期依赖关系，通过识别和学习节假日和特殊事件对货量的影响，组合模型能够在未来预测中考虑这些因素，增强对周末、节假日和其他特殊情况的适应性，提供更符合实际的预测结果。

3. 组合模型的针对性强。模型不仅关注每日货量变化，还深入分析每小时的货量波动，捕捉更细微的时间序列特征。针对每一个分拣中心，都进行分小时的货量分析后再进行预测。分小时的预测结果可以帮助分拣中心合理安排人员排班，避免人力资源浪费或不足，提高工作效率。

4. 针对每一个分拣中心，单独进行了特征分析后再进行预测，针对性强。不同分拣中心的业务量和工作模式可能存在差异，通过对每个分拣中心的数据进行单独分析和处理，确保模型能够针对性地适应各自的特点，提高预测的准确性。

5. 模型的鲁棒性强。单一模型在某些情况下可能存在局限性，而组合模型通过优势互补，减少了这些局限性对预测结果的影响。通过将两个不同类型的模型结合，ARIMA-LSTM 组合模型能够在不同情况下提供稳定的预测结果，提高模型的鲁棒性。

6.2 模型的缺点

1.LSTM 作为深度学习模型，需要大量的计算资源和数据，并且训练时间较长。当与 ARIMA 结合使用时，特别是处理大规模数据集时，整个混合模型的训练和预测过程可能会更加复杂和耗时。

2. 每次运行模型的结果存在一定随机误差。物流数据中不可避免地包含各种随机因

素，如突发订单、节假日效应、天气影响等。这些随机性因素会引入噪声，尽管对异常值进行了预处理，但一些无法完全消除的异常数据仍会对模型产生随机影响，导致预测结果出现误差，影响模型的预测准确性。

3.LSTM 模型具有很强的拟合能力，如果数据样本量不足或者数据噪声较大，模型容易出现过拟合现象，导致对新数据的泛化能力下降。

4. 缺少了灵敏度分析。ARIMA 模型的自回归项 (p)、差分项 (d) 和移动平均项 (q)，LSTM 模型的学习率、隐藏单元数、层数等参数对预测结果有显著影响，并且在组合模型中，ARIMA 和 LSTM 模型的权重分配对最终预测结果有直接影响，但未对权重变化对预测结果的敏感性进行系统分析。

6.3 模型的推广

ARIMA-LSTM 能推广到金融市场预测、气象预报、供应链管理和医疗健康等多个领域。通过时间序列预测，该模型能够帮助各行业优化资源配置、提高运营效率和服务水平，为科学决策提供可靠的数据支持。

参考文献

[1] Chun-Hua Chien and Amy J. C. Trappey. On the application of arima and lstm to predict order demand based on short lead time and on-time delivery requirements. Processes, 9(7):1157, 2021.

[2] Saroj. Review: Study on simple k mean and modified k mean clustering technique. 2016.

[3] Hugo Tsugunobu Yoshida Yoshizaki, Celso Mitsuo Hino, and Carlos Eduardo Cugnasca. Deep learning and statistical models for forecasting transportation demand: A case study of multiple distribution centers. Logistics, 7(4):86, 2023.

[4] 吕志燕 and 王培进. 基于 arima-lstm 的公路交通运输量预测. 公路与水路运输; 自动化技术, 2023.

[5] 张瀚文, 于长程, 李依婷, and 宁世雄. 基于 lstm 和 arima 的负荷预测对比分析. 智能电网（汉斯）, 2023.

[6] 杨艳, 黄晴, 龙思, 潘自翔, and 欧阳瑞祥. 基于 arima-lstm 的货运量组合预测方法研究. 公路与水路运输, 2022.

[7] 王代君, 李明, and 鹿守山. 基于 bayes-arima 的景区公路短时交通流量预测. 公路与水路运输, 2024.

[8] 龙宇, 许浩然, 余华云, 何勇, and 徐红牛. 基于 arima-lstm-xgboost 组合模型的铁路货运量预测. 科学技术与工程, 23(25):10879–10886, 2023.