分类号_____ 密级 _____

UDC _____ 编号 10741

硕士学位论文

论文题目多模态数据驱动的组合预测方法研究及应用

研究生姓名：于婷

指导教师姓名、职称：孟生旺 教授

学科、专业名称：统计学 数理统计学

研究方向：复杂数据分析

提交日期：2022 年 5 月 30 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：于婷 签字日期：2022.5.30

导师签名： 签字日期：2022.5.30

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意(选择”同意”/”不同意”)以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学”中国学术期刊(光盘版)电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：于婷 签字日期：2022.5.30

导师签名： 签字日期：2022.5.30

Research on multimodal data-driven combined forecasting method and its application

Candidate : Yu Ting

Supervisor : Meng Shengwang

摘 要

为了对时间序列数据进行更高精度的预测，本文基于”分解-集成”框架，提出了组合预测新方法，分别建立一次分解、二次分解、基于网络搜索信息的预测模型。

本文首先提出了一种基于 EEMD-PR/PSO-LSSVR-PM 的组合预测模型。首先，采用集合经验模态分解(EEMD)将数据分解为多个不同频率的本征模态函数(IMFs)，以降低数据的复杂性。然后利用粒子群最小二乘支持向量回归(POS-LSSVR)来分别预测 IMFs，利用一元多项式(PR)来预测带有趋势的残差项，最终对各个序列利用感知机模型(PM)进行非线性集成，得到最终预测结果。在实证分析中，以 2019 年我国港口集装箱吞吐量排名前十的大规模港口为研究对象，先运用 K-Means 聚类按数据特征将港口分为 3 类，从每一类中分别选取广州、营口、上海为三大代表性港口。利用本文所提出的组合预测方法进行实证预测。

其次，在一次分解的基础上，提出了一种二次分解-集成预测模型。在一次分解降低原始数据的复杂度的基础上，进一步挖掘数据潜在特征，并且通过重构子序列避免预测时的误差累积。第一步，用集合经验模态分解(EEMD)将原始机场客流量分解，将得到的子序列重构，得到高、中、低频序列；第二步，高、中频序列由于其变化波动较大、频率较快，采取变分模态分解(VMD)方法对其进一步分解，使其均被分解为复杂度较低，且更易于预测的子序列；第三步，采用布谷鸟搜索算法优化 BP 神经网络(CS-BP)模型预测所有子序列，并采用试错法自适应的确定神经网络模型最佳滞后期；第四步，分别将高频、中频子序列的预测值采用 CS-BP 模型进行集成，得到高频、中频的预测值；最后，将所有高、中、低频的预测值采用 CS-BP 模型汇总集成为最终预测值。

最后，在”分解-集成“框架下提出了一种基于网络搜索信息的组合预测新方法。首先，采用平均影响值和时差相关分析法对机场旅客吞吐量相关的网络搜索关键词进行筛选，利用每个关键词搜索量与原始航空客流数据的相关程度确定最佳滞后期，进而合成综合搜索指数。其次，利用 ICEEMDAN 方法分别将机场旅客吞吐量和综合搜索指数分解为若干子模态序列，并依据子序列的样本熵值重构为高、中、低频序列。以搜索指数中的不同频率成分作为辅助输入信息，分别对机场旅客吞吐量的高频和中频序列采用麻雀搜索算法优化的 BP 神经网络

(SSA-BP)模型进行预测，而低频序列采用自回归分布滞后模型进行预测，最后将不同频率序列预测值用 SSA-BP 进行综合集成得到最终的预测值。

利用文章所提出的组合预测方法进行实证研究预测，结果表明，本文提出的方法在在港口集装箱吞吐量预测和机场客流预测中均有较高的预测精度和鲁棒性。

关键词：集装箱吞吐量 机场客流预测 分解-集成 二次分解 网络搜索信息 布谷鸟搜索算法 麻雀搜索算法

Abstract

In order to predict time series data with higher accuracy, based on the “decomposition-integration” framework, this thesis proposes a new method of combined forecasting, and establishes a forecasting model based on primary decomposition, secondary decomposition, and network search information. Empirically, it is found that the combined forecasting method can significantly improve the forecasting accuracy and show better robustness.

In this thesis, a combined prediction model based on EEMD-PR/PSO-LSSVR-PM is proposed. First, ensemble empirical mode decomposition (EEMD) is used to decompose the data into multiple eigenmode functions (IMFs) with different frequencies to reduce the complexity of the data. Then use Particle Swarm Least Squares Support Vector Regression (POS-LSSVR) to predict IMFs separately, use Univariate Polynomial (PR) to predict residuals with trends, and finally use Perceptron Model (PM) for each sequence to perform nonlinear Integrate to get the final prediction result. In the empirical analysis, taking the top ten large-scale ports in my country's port container throughput in 2019 as the research object, the K-Means clustering was used to classify the ports into three categories according to data characteristics, and from each category, Guangzhou, Yingkou and Shanghai are the three representative ports. Empirical forecasting is carried out using the combined forecasting method proposed in this thesis.

A quadratic decomposition-integrated prediction model is established. On the basis of reducing the complexity of the original data by the primary decomposition, the potential features of the data are further mined, and the accumulation of errors in prediction is avoided by

reconstructing the subsequences. In the first step, the original airport passenger flow is decomposed by Ensemble Empirical Mode Decomposition (EEMD), and the obtained subsequences are reconstructed to obtain high, medium and low frequency sequences; in the second step, the high and medium frequency sequences fluctuate greatly due to their changes. , the frequency is faster, and the variational modal decomposition (VMD) method is used to further decompose it, so that it can be decomposed into subsequences with lower complexity and easier to predict; the third step is to use the cuckoo search algorithm to optimize the BP The neural network (CS-BP) model predicts all the subsequences, and uses the trial-and-error method to adaptively determine the optimal lag period of the neural network model; the fourth step is to use the CS-BP model for the predicted values of the high-frequency and intermediate-frequency subsequences. Integrate to get the predicted values of high frequency and medium frequency; finally, use the CS-BP model to aggregate and integrate all the predicted values of high, medium and low frequency into the final predicted value.

This thesis proposes another new method of “decomposition-int egration” combination prediction based on network search informatio n. Firstly, the average impact value and time difference correlation analysis method are used to filter the network search keywords rela ted to the airport passenger throughput, and the optimal lag period is determined by the correlation between the search volume of each keyword and the original aviation passenger flow data, and then a comprehensive search is synthesized. index. Secondly, the airport p assenger throughput and comprehensive search index are decompose d into several sub-modal sequences by the ICEEMDAN method, an d reconstructed into high, medium and low frequency sequences acc ording to the sample entropy values of the subsequences. Using the

different frequency components in the search index as auxiliary in put information, the high-frequency and intermediate-frequency seque nces of the airport passenger throughput are predicted by the BP ne ural network (SSA-BP) model optimized by the sparrow search algo rithm, while the low-frequency sequences are predicted by the autor egressive distribution. The lag model is used for prediction, and fin ally the prediction values of different frequency series are integrated with SSA-BP to obtain the final prediction value.

The combined forecasting method proposed in this thesis is used to conduct empirical research forecasting. The results show that the method proposed in this thesis has high forecasting accuracy and robustness in both port container throughput forecasting and airport passenger flow forecasting.

Keywords： Container Throughput;Airport Passenger Flow Forecast; Decomposition-Integration;QuadraticDdecomposition;Network Search Information;Cuckoo Search Algorithm;Aparrow Aearch Algorithm;

目 录

1 绪论

1.1 研究背景及意义

在我国的发展过程中，铁路、公路、水路、航空、管道等五大交通系统的发展，客观上需要构建与之相适应的一体化运输系统。在这一新的形势下，我国的交通运输业面临着许多新的问题和新的任务。结合我国目前的运输发展状况，研究适合我国交通运输行业预测模式是一项新的研究课题。航空客流预测是综合运输客流量预测中的一个重要内容，港口集装箱吞吐量是反映港口生产经营活动成果的重要指标，是当今社会国际运输的主要方式之一，这些都是有待进一步研究的问题。

现如今进出口贸易不断发展，国内外物资交流的同时为港口集装箱的规划与布局带来巨大挑战，长期监测港口集装箱吞吐量并对其未来趋势进行相关的预测研究，是确定港口发展规模，规划港口总体布局的关键。如果对港口集装箱吞吐量的预测出现重大偏差，可能会给港口的发展带来无法估量的经济损失，甚至会影响全国经济的发展。因此，对我国重要港口的集装箱吞吐量进行精准预测，可以为政府和相关企业提供准确的决策参考，这对国家的长期发展也具有重大意义。

同时，随着经济一体化的深入，各大一线、新一线城市的人口密度都在不断的增加，机场的客流量也在不断的增加，这也是一个国家和地区的经济和社会发展的重要因素。首先，机场的资源相对来说比较有限，值机、行李托运、安检、候机点的安排、紧急情况下的应变，都需要对未来的客流量进行空间分布，根据现有的客流特征进行合理的调整，从而最大限度地发挥机场的资源利用率，节省运营费用，减少排队等待时间，让乘客得到更满意的用户体验。在对预测方法的不断探索中，管理者和有关学者不断改进和创新，从而找到能够满足需求的预测方法，促进预报技术的持续发展，科学的决策常常是基于科学的预测。通过对未来不确定性的定量、定性的分析，为决策者提供科学依据。机场旅客吞吐量预测是机场系统规划的先决条件，因此，必须通过合理的预测和分析规划出未来的机场设施和设备，并对其进行合理的安排。机场旅客流量预测：对未来的运输网络进行合理的规划，能有效地适应目前复杂的运输网络与经济发展模式，推动我国

国民经济的健康发展。机场的客流量预测对于区域内的机场建设和机场发展具有一定的参考价值。机场的客流量是影响其生存与发展的重要因素。而机场的服务，除了要提高员工的素质之外，还需要事先了解到每个航班的客流量，才能进行合理的调配。此外，对机场的客流量进行预报，不仅可以提升机场的效率，而且可以降低延误率等意外事件。

1.2 研究现状

1.2.1 国内外研究现状

(1)传统时间序列预测模型

基于单一传统时间序列的预测方法，是预测航空客流、港口集装箱吞吐量最早的方法，常用到的指数平滑法、ARIMA、SARIMA、灰色模型、多元线性回归等，到目前也是主流预测方法之一。刘斌等[1](2003)建立回归模型来预测港口集装箱吞吐量；张丽等[2](2006)通过对比 SARIMA 模型与基于加法模型的 Holt-Winters(HW)指数平滑法对中国航空旅客运输量的预测，提出 SARIMA 在建立区间预测时更可取，而 HW 指数平滑法在点预测时精度更高；姚晏斌等[3](2006)通过对影响机场旅客吞吐量的主要因素进行灰色关联分析，建立 GM(1,2)预测模型，并且根据首都机场旅客吞吐量的实例进行预测；Ching 等[4](2009)通过比较 Holt-Winters(HW)方法、SARIMA 模型和 GM(1,1)来预测台湾的每月入境航空旅行人数，提出 SARIMA 模型是预测台湾入境航空旅客人数的最佳模型；李明捷等[5](2009)运用三次指数平滑法对首都国际机场旅客吞吐量进行了预测；Erma 等[6](2010)结合机票价格影响、服务水平影响等因素建立多元线性回归来预测机场的年旅客吞吐量；黄邦菊等[7](2013)以西南某枢纽机场为研究对象，考虑该地区的经济、人口因素等 9 项相关数据作为因变量，建立多元线性回归分析预测模型，最后提出该地区建立第二机场时有必要的；Tsui 等[8](2014)用 SARIMA 和 ARIMAX 模型对香港机场客流量进行预测，并表明香港机场未来客运量将继续以不同幅度增长；Kim 等[9](2016)利用大数据建立仁川国际机场短期航空旅客需求预测模型，以该机场的月客运量作为预测模型的因变量，互联网搜索数据为自变量，建立回归分析。王婷婷等[10](2017)运用灰色马尔科夫模型对龙洞堡机场旅客吞吐量进行预测，发现该方法对旅客吞吐量进行预测研究是切实可行的；杨梦达等[11](2019)进一步进行改进，通过对剔除季节因素后的上海虹桥机场旅客吞吐

量进行预测，模型更为准确，可以进行短期预测。

(2)机器学习预测模型

机器学习的预测模型主要包括 BP 神经网络、极限学习机、支持向量机、长短期记忆神经网络等；还有是通过智能优化算法对神经网络的参数进行优化，提高神经网络模型的预测精度，如遗传优化算法优化极限学习机，花授粉算法(FPA)、人工蜂群算法(ABC)等优化 BP 神经网络。刘长俭等[12](2007)建立 BP 神经网络开始预测集装箱吞吐量，同时也有学者将 RBF 神经网络[13-14]，Elman 神经网络[15]、支持向量回归(Support Vector Regression,SVR)[16]等应用到了港口集装箱吞吐量预测的研究中，肖海波等[17](2005)利用 BP 神经网络对成都双流国际机场的旅客吞吐量进行预测，而且有更好的预测精度；冯兴杰等[18](2005)以成都国际机场作为研究对象，发现采用支持向量回归方法进行机场旅客吞吐量预测是可行的；Yan 等[19](2009)为了提高航空客流的预测能力，引入支持向量回归；廖洪一等[20](2015)通过遗传算法对极限学习机的参数进行寻优，建立成都双流国际机场旅客吞吐量预测模型；王子位等[21](2018)基于长短期记忆网络构建了民航流量的预测模型，从而实现对短期民航流量的高精度预测；李洁等[22](2018)基于多时间尺度的 RNN 预测模型分别对旅客出行情况的短期性时间依赖、周期性时间依赖和长期性时间依赖进行建模，从而对未来时段各个机场的客流量情况进行预测；Sun 等[23](2019)提出了一种基于 MIV 的非线性向量自回归神经网络(NVARNN)的航空客流预测方法，利用北京国际机场客流量说明和验证该方法的有效性；Korkmaz 等[24](2021)首次应用五种不同的优化算法开发了不同的预测模型，即花授粉算法(FPA)、人工蜂群算法(ABC)、乌鸦搜索算法(CSA)、磷虾群算法(KH)和蝴蝶优化算法(BOA)来估计土耳其的航空运输需求；Chan 等[25](2021)使用神经 Granger 因果关系模型可以识别出谷歌趋势查询预测因子，从而提高樟宜机场到达旅客的预测精度。

(3)混合模型

混合预测模型主要分为两类，一类是集成两个或两个以上的方法，集合多种模型的优点提高模型的预测精度，具体是多种传统时间序列预测模型的结果进行集成、多种机器学习的预测结果进行集成或传统时间序列预测于机器学习预测结果进行集成，集成的方法可以选择简单平均、加权平均或是用神经网络进行集成。另一类是分解-集成模型，即就是将数据进行分解，得到若干个子序列，不同子

序列逐个进行预测，最后将预测值进行集成，得到最终预测结果。进一步，在机场旅客吞吐量预测时，数据的选择方面，已经不局限于机场旅客吞吐量，或是传统线性回归选选择的影响因素，如 GDP、人口等因素，随着互联网时代的发展，越来越多的人通过电脑手机等设备来获取信息，搜索引擎的搜索量可以在一定程度上反应人们对相关事物的关注度以及对某些事物的需求，为了弥补传统数据的不足，越来越多的学者把互联网搜索查询量引入模型，加入到分解集成模型中。

赵尚威等[26](2018)，蒋惠园等[27](2020)通过将不同的单一预测方法进行组合预测集装箱吞吐量，同时采用不同方法确认权重，得到了更好的预测精度。冯宏祥等[28](2021)提出 EMD-SVR 模型来预测上海港集装箱吞吐量，该模型用经验模态分解(Empirical Mode Decomposition，EMD)方法将原始数据分为数个频率不同的子列，对各子列用 SVR 进行预测，最后将各子列的预测结果进行集成，得到最终预测结果。屈拓等[29](2012)用灰色模型、BP 神经网络两种模型进行组合对成都双流国际机场旅客吞吐量进行预测，解决了单一预测模型的缺点，减小了预测误差；Gang 等[30](2014)提出了混合方法用于机场航空旅客的短期预测，将序列季节分解，子序列用最小二长支持向量回归预测，最后集成输出最后预测结果；梁小珍等[31](2017)把剔除噪声的原始的数据分别用 ARIMA、SVM、HW 指数平滑法进行预测，最后通过加权平均(MA)把三个单一模型的预测结果进行集成得到最终的预测结果，得到较高预测精度；刘夏等[32](2018)采用 HW 指数平滑法、差分自回归移动平均模型(ARIMA)和一元线性回归三种模型分别对三亚机场客流量进行了预测，将预测精度较好的 HW 指数平滑模型和 ARIMA 模型组合加权求得最终的预测值，相比单一模型，提高了预测精度。Jin 等[33](2019)为了更精确地捕捉航空客流数据特征，将分解后的子序列分别采用传统 ARMA 模型与神经网络 KELM 进行预测，最后进行非线性集成得到最终结果；梁小珍等[34](2021)把网络搜索指数航空客流量进行结合，建立两阶段分解集成预测模型。

1.2.2 文献评述

通过对现有国内外的预测方法进行研究和梳理，我们得到如下关于该领域的结论及启示：

现如今，“先分解后集成”的思想是时间序列预测较为前沿和领先的技术。分析相关文献，关于时间序列的预测方法可大致分为 3 类：传统时间序列预测模型、机器学习模型、混合预测模型。

(1)传统的时间序列预测模型理论成熟，主要是对规律性较强、线性特征明显的数据有较高的预测精度，但机场旅客吞吐量是一组高波动、非线性、非平稳的时间序列数据，使用多元线性回归模型和 ARIMA 等传统的时间序列的预测方法均未能很好的处理模型中非线性、高波动等问题，而灰色模型只适用于小样本，并且它的预测精度主要依赖于灰度参数的设定。

(2)机器学习模型较为灵活，不会对样本有太多的限制，也没有过多的假设；但依旧存在一些无法规避的缺陷，如参数敏感性、局部最优及过度拟合等。

(3)混合模型模型则是根据数据的特征，结合多种方法，这样能充分发挥每个模型的特点，形成优势互补，在时间预测领域具有独特的优势。

在时间序列预测领域，以“分解集成”思想构建模型，成为现如较为主流的预测方法，通过将时间序列数据分解为复杂度较低，易于分析与预测的子序列，降低了预测难度，提高模型的分析与预测性能。同时，“数据特征”在时间序列预测中起着非常关键作用，在数据的预测方面，不论是新的预测方法提出，框架的完善，均围绕研究样本的数据特征展开。

因此，数据特征是时间序列数据预测创新的一个基本出发点，充分研究样本的数据特征，构建与其特征相适应的预测模型，使模型具有普适性，进行二次分解充分挖掘数据特征，降低非平稳非线性数据的预测难度，结合网络搜索信息，拓展预测时间序列的辅助信息，为预测研究提供一个新的创新思路。

1.3 技术路线及文章安排

本文以港口集装箱吞吐量、机场客流预测为研究内容，将广州港、营口港、上海港，广州白云国际机场、昆明长水国际机场、青岛胶东国际机场、西安咸阳国际机场、成都双流国际机场的客运量作为研究对象。对数据进行预处理，掌握航空客流数据的特征，针对其非平稳、非线性的特性，我们的目的就是想办法降低其数据的复杂性，选取适合其特征预测方法构建模型，从而提高预测精度。通过不同分解技术将时间序列数据分解为具有不同特征的多个序列，并且通过度量子序列的数据特征，进一步进行分析与预测。本文从技术路线出发，分为三个部分：(1)对于复杂数据，基于”先分解后集成”的思路对数据进行分解，针对数据特征，采用不同的方法进行预测，最后将预测数据进行非线性集成，采用”分而治之”的策略提高预测精度。(2)一些高频的数据波动较快，单一的分解方法不能充

分提取数据的相关特征，作为对一次分解的进一步提升，文章通过对数据进行二次分解，降低高频序列的复杂度，降低建模难度，提高个个序列的预测精度，接着将各子序列的预测结果通过非线性集成的方法得到最终的预测结果。对比其他模型，用不同的评测标准来检验文章所提出模型的精度和鲁棒性。文章提出的二次分解方法，弥补了单一模型的缺点，整合相同特征的数据，深度挖掘数据的特征，提高了航空客流数据预测的精度。(3)为跟随大数据时代的进步与发展，文章提出结合网络搜索信息建立模型来预测航空客流。在预测航空客流数据时，采用多元数据可以提高模型的预测精度，文章提出的网络搜索信息不仅从机场自身角度出发，同时添加机场所在地旅游相关信息，对网络搜索关键词进行筛选与整合，丰富预测时辅助信息的多样性。

本文总共由五个章节构成：

第一章绪论，阐述文章港口集装箱吞吐量、机场旅客吞吐量的研究背景与意义，针对国内外在这两方面预测领域的文献所做研究进行归纳总结，介绍文章内容安排以及文章创新点；

第二章对文章中所用到的方法逐一进行介绍，包括分解方法、重构方法、预测方法以及建立网络搜索信息指标所用到的方法；

第三章为第一个实证研究：基于分解集成框架下的港口集装箱吞吐量预测，介绍如何选取预测数据，对其进行分解预测，集成最终预测结果，分析比较模型；

第四章为第二个实证研究：基于二次分解集成框架下的航空客流预测，将建模过程中数据预处理、进行分解预测、分析比较相关模型优劣，得出相应结论；

第五章为第三个实证研究：分解集成框架下基于网络搜索信息的航空客流预测，该部分介绍在建立网络搜索指标体系时，关键字的选取、合成，将网络搜索信息与时间序列数据相结合建立分解-集成预测模型并分析结果。

第六章总结，将着重写出本文的研究结论，同时列出本文的不足之处。

1.4 论文创新点

本文主要在分解集成的思想下建立基于数据驱动的时间序列预测，在本文的应用中，主要创新点如下：

（1）对数据分解后基于不同特征子序列选取不同的预测模型，”分而治之”的策略从逻辑上保证了本文模型能够提高预测精度；

（2）提出二次分解-重构模型，根据样本熵值度量数据特征，对数据分类预测，提高模型精度。

（3）将网络搜索关键词的筛选范围从机场自身信息扩充到了与当地旅游相关的信息，从而进一步拓宽了网络搜索关键词的维度，丰富了预测的信息源。

（4）综合使用平均影响值和时差相关分析法优选搜索关键词，并基于搜索量信息与机场客流量的相关程度，通过加权平均构造出综合搜索指数；

（5）对子序列预测时采用智能优化算法粒子群优化算法优化支持向量机参数、麻雀搜索算法、布谷鸟算法来优化 BP 神经网络的权值和阈值，使得神经网络的预测性能得到进一步提高。

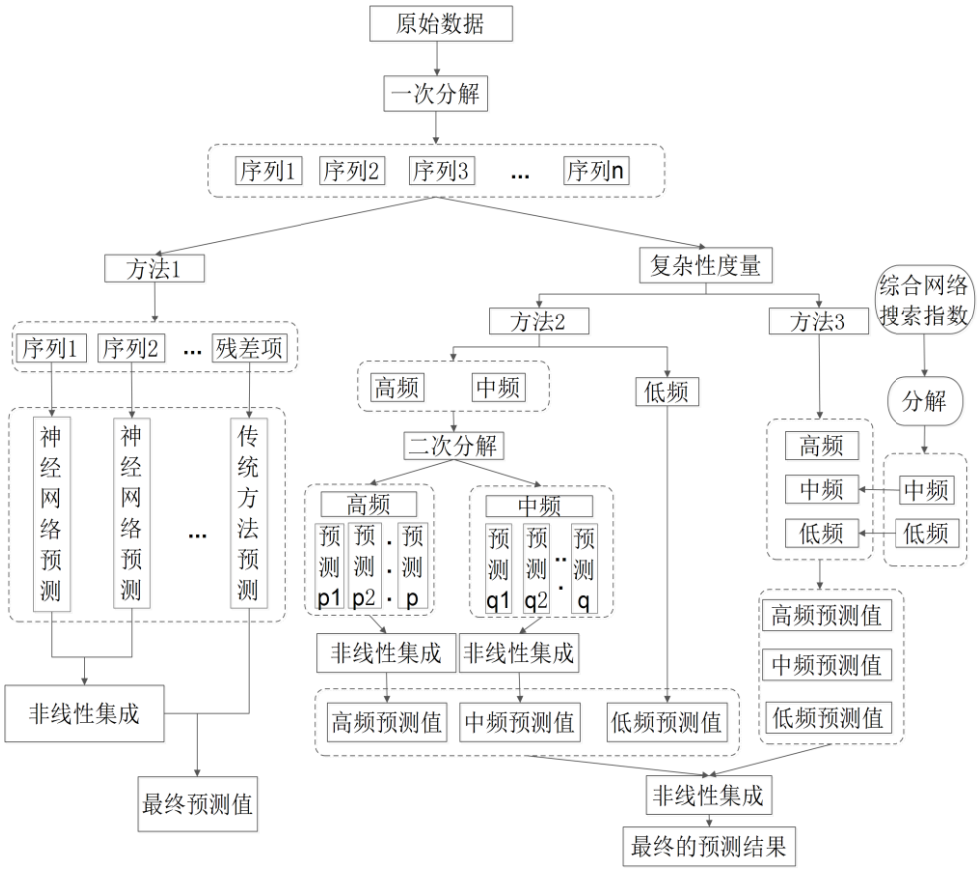

图1.1 技术路线图

2 研究方法

2.1 数据分解方法

2.1.1 集合经验模态分解(EEMD)

EMD[35]是一种自适应的信号分解算法，它根据数据本身的特性将其划分成具有多种特征尺度的子序列，并将其称为本征模态函数(IMF)和残差(Res)。这种方法可以很好的处理非平稳和非线性的时序。但是，EMD 分解法有许多缺点，如模态混叠、端点效应。改进的集经验模式分解(EEMD)[36]法，将白噪声信号添加到原始信号中，从而改变极值点的分布，解决了 EMD 分解时的模态混叠问题。EEMD 的具体步骤是：

(1) 将原始数据 $x(t)$ 加入正态分布的白噪声 $w(t)$，得到 $y(t)=x(t)+w(t)$；

(2) 对信号 $y(t)$ 进行 EMD 分解得到 $y(t)=\sum_{i=1}^{n}C_i(t)+r_n(t)$，$C_i(t)$ 是 EMD 分解得到的第 i 个 IMF，$r_n(t)$ 是分解得到的序列残差分量；

(3) 重复步骤(1)和(2) M 次，每次添加不同幅值的白噪声，获得 M 组 IMF，计算其均值，得 EEMD 得 IMF 分量及残差分量：

$$\bar{C}_i(t)=\sum_{m=1}^{M}C_{i,m}(t)\Big/M \tag{2-1}$$

$$\bar{r}_n(t)=\sum_{m=1}^{M}r_{n,m}(t)\Big/M \tag{2-2}$$

2.1.2 改进的自适应噪声完备集合经验模态分解(ICEEMDAN)

完备集合经验模态分解(CEEMD)[37]是加入一对正负白噪声序列克服 EMD 方法的缺陷，自适应噪声完备集合经验模态分解(CEEMDAN)[38]是在 CEEMD 方法的基础上，加入自适应的白噪声，使得模型重构误差趋近于零且解决分解得到的虚假成分问题。ICEEMDAN 则是 CEEMDAN 方法的进一步改进，该方法降低了 IMF 中的残余噪声，还解决了分解早期阶段的虚假成分和模态混叠问题，目前 CEEMDAN、ICEEMDAN 等数据分解方法已经广泛的应用于人民币汇率预测、能源价格预测等领域[39-40]。ICEEMDAN 的具体分解过程可以描述如下：

假设 $w^{(i)}$ 表示要添加的第 i 个白噪声，$E_k\left(w^{(i)}\right)$ 表示对白噪声 $w^{(i)}$ 进行 EMD

分解后得到的第 k 个 IMF 分量，β_{k-1} 表示第 k 次添加的白噪声序列标准差。首先将第一个白噪声分量 $E_k\left(w^{(i)}\right)$ 加到原始信号 x 上，得到如下信号 $x^{(i)}$：

$$x^{(i)} = x + \beta_0 E_1\left(w^{(i)}\right) \tag{2-3}$$

用 EMD 分解得到 $x^{(i)}$ 的局部均值，取其平均值得到第一个残差 $r_1 = \frac{1}{T}\sum_{i=1}^{T} M\left(x^{(i)}\right)$。其中，$M(\cdot)$ 表示信号的局部均值，T 为添加的白噪声序列个数。然后，原始序列 x 的第一个 IMF 值可以通过 $c_1 = x - r_1$ 来计算。

第二个模态分量值(IMF2)可通过以下公式计算：

$$c_2 = r_1 - r_2 \tag{2-4}$$

其中 $r_2 = \frac{1}{T}\sum_{i=1}^{T} M\left(r_1 + \beta_1 E_2\left(w^{(i)}\right)\right)$。

同样，根据 $c_k = r_{k-1} - r_k$ 计算第 k 个 IMF 值，其中 $r_k = \frac{1}{T}\sum_{i=1}^{T} M\left(r_{k-1} + \beta_{k-1} E_k\left(w^{(i)}\right)\right)$，至此，原始信号被分解为：

$$x = \sum_{n=1}^{k} c_n + r_k \tag{2-5}$$

2.1.3 变分模态分解(VMD)

VMD 是由 Konstantin[41]等(2014)提出了一种解决输入信号中噪声问题的分解技术，可以自适应地确定频带并估计相适应的模态。通过合理地设置预设尺度数 K 的值，变分模态分解便可以有效地抑制在 EMD 分解中经常出现的模态混叠现象。

Konstantin 等的实验结果显示，在对噪声的处理上，VMD 的性能明显优于 EMD，具有更显著的鲁棒性。与经验模态分解(EMD)相比，VMD 在对非平稳信号进行处理时，可以更显著地抑制噪声的影响。VMD 分解的主要目的是将实值的输入信号分解为离散的模态分量序列，同时这些分量序列在再现输入时可以具有特定的稀疏性。VMD 分解的主要思路是通过求解变分问题，确定每个模态的中心频率 ω_k 和带宽 α。将原始序列 f 分解为一系列有限带宽模态函数 $\{u_k(t)\}$，$k = 1, 2, \ldots K$，其主要过程如下：

(1)原始的约束变分问题可以建立变分算式来让各本征模态函数的带宽之和最小：

$$\begin{aligned}&\min_{\{u_k\},\{\omega_k\}}\left\{\sum_k\left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-jw_kt}\right\|_2^2\right\}\\&s.t.\sum_k u_k=f\end{aligned}\tag{2-6}$$

其中，$\{u_k\}=\{u_1,u_2,\cdots,u_k\}$ 表示所有模态分量；$\{\omega_k\}=\{\omega_1,\omega_2,\cdots,\omega_k\}$ 表示各子模态的中心频率；$\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)$ 为模态函数 $u_k(t)$ 采用 Hilbert 变换得到的解析信号，$\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-jw_kt}$ 为每个模态频谱调制得到相应基频带，$\left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-jw_kt}\right\|^2$ 为每个模态的频带宽度。

(2)采用二次惩罚因子 α 和拉格朗日乘子 λ_t，得到增广拉格朗日表达式，将约束性变分问题变为非约束性变分问题：

$$L(\{u_k\},\{\omega_k\},\lambda)=\alpha\sum_k\left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-jw_kt}\right\|_2^2+\left\|f(t)-\sum_k u_k(t)\right\|_2^2+\left\langle\lambda(t),f(t)-\sum_k u_k(t)\right\rangle\tag{2-7}$$

(3)使用交替方向乘子法对这个问题进行迭代收敛，把时域上的各模态分量 $u_k(t)$、$f(t)$、$\lambda(t)$ 进行傅里叶变换转化到频域 $\hat{u}_k(\omega)$、$\hat{f}(\omega)$、$\hat{\lambda}(\omega)$，于功率谱找到中心来确定其中心频率 $\hat{\omega}_k$。交替更新 $\hat{u}_k^{n+1}$、ω_k^{n+1}、λ^{n+1}：

$$\hat{u}_k^{n+1}(\omega)=\frac{\hat{f}(\omega)-\sum_{i\neq k}\hat{u}_i(\omega)+\frac{\hat{\lambda}(\omega)}{2}}{1+2\alpha(\omega-\omega_k)^2}\tag{2-8}$$

$$\omega_k^{n+1}=\frac{\int_0^\infty\omega\left|\hat{u}_i(\omega)\right|^2d\omega}{\int_0^\infty\left|\hat{u}_i(\omega)\right|^2d\omega}\tag{2-9}$$

$$\hat{\lambda}^{n+1}(\omega)\leftarrow\hat{\lambda}^n(\omega)+\tau\left(\hat{f}(\omega)-\sum_k\hat{u}_k^{n+1}(\omega)\right)\tag{2-10}$$

收敛条件为：

$$\sum_{k=1}^K\left\|\hat{u}_k^{n+1}-\hat{u}_k^n\right\|_2^2\Big/\left\|\hat{u}_k^n\right\|_2^2<\varepsilon\tag{2-11}$$

当满足收敛条件时，停止迭代，得出 $\hat{u}_k$ 和中心频率 ω_k，最后对 $\hat{u}_k(\omega)$ 逆变换得 $u_k(t)$ 即为模态分量。

2.2 样本熵(SE)

样本熵(Sample Entropy,SE)[42]是度量时间序列复杂程度的一种方法。样本熵值越大，说明该序列的复杂程度越高，反之亦然。本文使用样本熵来度量分解后各子序列的复杂度，为序列的重构提供依据。具体的样本熵计算过程如下：

对时间序列$\{y_i\}$ $i=1,2,...,N$，取相邻m个y_i组成新的序列

$$Y_i=[y_i,y_{i+1},...,y_{i+m-1}],i=1,2,...,N-m+1 \tag{2-12}$$

按下式计算距离

$$D_m\left(Y_i,Y_j\right)=\max\left\{\left|y_{i+l}-y_{j+l}\right|\right\},l=0,1,2,...,m-1 \tag{2-13}$$

其中$j=1,2,...,N-m$且$i\neq j$。

计算Y_i与Y_j的距离小于给定阈值r的序列个数B_i在总序列个数中的占比：

$$B_i^m(r)=\frac{B_i}{N-m} \tag{2-14}$$

计算$B_i^m(r)$的均值$B^m(r)$

$$B^m(r)=\frac{1}{N-m+1}\sum_{i=1}^{N-m+1}B_i^m(r) \tag{2-15}$$

将维数从m增加到$m+1$，重复上述步骤，得到$B^{m+1}(r)$.

计算样本熵值

$$SE(m,r)=\lim_{N\to\infty}\left\{-\ln\frac{B^{m+1}(r)}{B^m(r)}\right\} \tag{2-16}$$

当N为有限值时，熵值的计算公式如下

$$SE(m,r,N)=-\ln\frac{B^{m+1}(r)}{B^m(r)} \tag{2-17}$$

2.3 支持向量机

支持向量机(Support Vector Machines,SVM)是由 Vapnik[43]提出的一种用于分类的算法，当其用于数值预测时，称为支持向量回归。

支持向量机对于非线性数据有很强的处理能力，其基本原理是找到一个非线性映射函数$\phi(x)$，将线性不可分的数据x从低维特征空间投影至高维特征空间，使得原本非线性的问题转化为线性问题。即在高维空间中，建立超平面：

$$f(x)=[\omega\cdot\phi(x)]+b,\phi:R^m\to F,\omega\in F \tag{2-18}$$

式中，m为空间维度，ω为权向量，b为偏差。这个函数近似问题与下面函

数的最小化问题等价：

$$R=\frac{1}{2}\|\omega\|^2+C\frac{1}{l}\sum_{i=1}^{l}|y_i-f(x_i)|_{\varepsilon} \tag{2-19}$$

其中，y_i是样本的实际输出，而$f(x_i)$是预期的输出值。l是样本数。$\|\omega\|$是加权向量范数，用于约束模型的结构容量，从而获得更好的泛化能力。C是确定性经验误差和正则项之间权衡的正则化常数。

我们通过支持向量机最小化目标函数来确定回归函数，引入如下目标函数：

$$\min\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}(\xi_i^*+\xi_i) \tag{2-20}$$

其中ξ_i^*和ξ_i为松弛变量，求最小化支持向量回归问题等价于解决以下最优问题：

$$\begin{cases}\min\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}(\xi_i^*+\xi_i)\\ y_i-\omega\cdot\phi(x_i)-b\le\varepsilon+\xi_i^*\\ \omega\cdot\phi(x_i)+b-y_i\le\varepsilon+\xi_i\\ \xi_i^*,\xi_i\ge 0,i=1,2,...,l\end{cases} \tag{2-21}$$

通过引入拉格朗日乘数α_i，α_i^*将支持向量回归函数转化为对偶问题进行解决，上式的解即为最优的判别函数：

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(X_i,X)+b \tag{2-22}$$

其中$K(X_i,X)$称为核函数，核函数的类型很多，常用的有两种核函数：线性核函数与 RBF 核函数。

线性核函数：$$K(X_i,X)=X_i^TX \tag{2-23}$$

RBF核函数：$$K(X_i,X)=\exp\left(-\|X_i-X\|^2/2\sigma^2\right) \tag{2-24}$$

本文中所用到的核函数为 RBF 核函数，σ^2是 RBF 内核函数参数，而线性内核函数没有参数。总之，惩罚参数C，RBF 内核函数参数σ^2的不同取值直接影响着最终的预测效果，本文引入了粒子群优化算法(PSO)来优化参数C和σ^2。

2.4 优化算法

在实际中使用 BP 神经网络模型时，模型中的初始权值ω与阈值b是随机给定的。然而，初始权值和阈值的不同取值会直接影响模型的预测精度。因此，本文引入布谷鸟搜索算法(Cuckoo Search,CS)、麻雀搜索算法(Sparrow Search Algorithm,SSA)对初始权值和阈值进行优化，进而提高模型的预测精度。运用

LSSVR 时往往难以达到我们对精度的要求，因为惩罚参数C和核参数σ^2的不确定性都会导致我们没办法得到最优的预测值，故而找到最适合模型的参数C和σ^2是提高预测精度的关键。利用粒子群算法对最小二乘支持向量机的惩罚参数C和核参数σ^2进行优化，尽可能寻找到全局最优解。

2.4.1 粒子群优化算法(PSO)

在实际运用 LSSVR 时往往难以达到我们对精度的要求，因为惩罚参数C和核参数σ^2的不确定性都会导致我们没办法得到最优的预测值，故而找到最适合模型的参数C和σ^2是提高预测精度的关键。本文利用粒子群算法对最小二乘支持向量机的惩罚参数C和核参数σ^2进行优化，尽可能寻找到全局最优解。

PSO 算法最初由 Kennedy 和 Eberhart[44]在 1995 年提出的，一种基于鸟类觅食行为的搜索算法。鸟类在捕食的过程中，每只鸟只需搜寻此时距离食物较近的鸟的周围，就可以用较短的时间获取到食物，人们从鸟类行为特征中得到启发并用于求解优化问题。

本文中具体用 PSO 算法来优化惩罚C和核参数σ^2的步骤如下：

步骤 1 初始化一群粒子。本文中种群规模$n=20$，包括初始化粒子的速度与位置，再设定学习因子$C_1=1.5$，$C_2=1.7$,最大迭代次数为$k=50$，惯性因子设置为$w=1.9$。

步骤 2 计算粒子适应度值。本文先用 5 折交叉验证对训练样本进行随机划分，将 5 次均方根误差的平均值定义为适应度函数，其值的好坏表示粒子的优劣。

步骤 3 迭代更新。 在每次迭代中，粒子每更新一次位置,就计算一次适应度值，据此来进行粒子的更新，粒子都通过两个位置进行更新：第一，更新实现个体历史最佳位置 Pbest；第二，更新种群中找到全局的最佳位置 Gbest。

步骤 4 更新粒子的位置与速度。更新公式为：

$$v_{ij}^{k+1}=wv_{ij}^{k}+c_1r_1\left(p_{ij}^{k}-x_{ij}^{k}\right)+c_2r_2\left(p_{ij}^{k}-x_{ij}^{k}\right) \tag{2-25}$$

$$x_{ij}^{k+1}=x_{ij}^{k}+v_{ij}^{k} \tag{2-26}$$

其中，w是惯性因子，$i=1,2,\ldots,N$，N是种群中粒子的总数；$j=1,2,\ldots,K$，K是搜索空间的维数；v_{ij}^{k}为粒子当前速度；x_{ij}^{k}为粒子当前位置；c_1、c_2是学习因子。

步骤 5 确定最优解。当迭代次数达到最大设定次数或当误差小于预设的误差精度，即停止迭代，输出结果。

2.4.2 布谷鸟搜索优化算法(CS)

Yang[45]等(2009)年提出的布谷鸟搜索算法是一个新的群智能优化算法。由于参数少，运行简单，执行容易，布谷鸟搜索算法已经引起了很多学者的关注，布谷鸟搜索算法的研究与应用与日俱增。它模拟了莱维飞行机制和布谷鸟寻找巢穴繁殖行为来寻求最优解。通过莱维飞行机制，布谷鸟搜索算法实现了短距离小步长和远距离大步长交替迭代。因此，这个算法有一个强的搜索能力。布谷鸟希望它的后代能被其他的鸟类孕育后代，它们通常将受精的鸟蛋产到其他鸟类的窝中，并且为了提高自己后代的存活率，从而移走宿主鸟类的蛋。布谷鸟搜索算法主要基于下面三个先决条件：

(1)每只布谷鸟每次只产一枚蛋，并且随机选择一个鸟窝进行孵化。

(2)具有优质蛋的鸟窝会被保留至下一代。

(3)鸟窝的总量一定，宿主鸟发现外来蛋的概率为 $p_a(p_a \in [0,1])$。宿主鸟发现外来蛋时，它会将外来蛋丢弃或者直接放弃该鸟窝重新盖窝。

基于上述三种先决条件，布谷鸟寻找宿主鸟窝的路径和位置更新公式为：

局部搜索：

$$X_i^{t+1} = x_i^t + \partial s \otimes H\left(P_a - \varepsilon\right) \otimes \left(x_j^t - x_k^t\right) \tag{2-27}$$

全局搜索：

$$X_i^{t+1} = x_i^t + \partial \otimes L\left(s,\lambda\right)，\ i = 1,2,\cdots n \tag{2-28}$$

其中，x_i^t 表示第 i 个鸟窝在第 t 代的鸟窝位置，$\partial > 0$ 为步长因子，来控制步长，且服从 $N\left(0,1\right)$ 分布，$\otimes$ 代表元素乘法，ε 为服从均匀分布的随机数，$H\left(\cdot\right)$ 为阶跃函数，$L\left(s,\lambda\right)$ 为莱维分布：$L\left(s,\lambda\right) \sim \dfrac{\lambda\Gamma\left(\lambda\right)\sin\left(\pi\lambda/2\right)}{\pi}\dfrac{1}{s^{\lambda}+1}\left(s \gg s_0 \gg 0\right)$，$\Gamma\left(\lambda\right)$ 表示标准伽马函数，λ 表示介于 $(1,3]$ 的分布因子，s 表示步长，s_0 则为最小步长。更新公式中有两个分支，局部搜索步骤主要用于局部细化，而主要的移动或探索由全局搜索步骤进行。

具体的算法步骤如下：

步骤 1 初始化设置鸟巢数量 n、p_a、最大迭代次数等参数，然后评估它们的目标值，从而找到当前的全局最佳。

步骤 2 计算 n 个鸟巢位置的目标函数值并进行比较，得到当前的最优函数值为全局最优。

步骤 2 鸟巢位置进行更新：

$$X_i^{t+1} = x_i^t + \partial \otimes L(\lambda) \tag{2-29}$$

步骤 3 从均匀分布 $[0,1]$ 中抽取一个随机数 ε。如果 $\varepsilon > p_a$，则更新 X_i^{t+1}，否则鸟巢位置不变，计算其目标函数值，更新最优解。

步骤 4 如果满足最大迭代次数或最小误差，则得到最优解。否则，返回步骤 2。

2.4.3 麻雀搜索优化算法(SSA)

SSA 是 Xue 和 Shen[46]于 2020 年提出的一种新型的群智能优化算法，该算法模拟一群麻雀觅食的过程。在算法规则中，将麻雀群体分为发现者和加入者，另外还有以麻雀为对象的捕食者。根据麻雀个体所对应的不同适应度分为发现者和加入者，发现者承担搜索食物丰富的区域，为种群提供觅食区域和方向，而加入者则通过发现者提供的信息来获取食物；二者可以相互转化但二者数量占比不变。在觅食过程中，部分麻雀发现捕食者时会发出警报，称为警戒者，当警报值大于安全值时，发现者会带领加入者转移。

该算法新颖，具有寻优能力强，收敛速度快的优点，具体的算法步骤如下：

步骤 1 初始化种群，迭代次数，初始化发现者和加入者比例。假设麻雀总数为 n 只，最大迭代次数为 $iterm_{\max}$，搜索空间的维度为 d。

步骤 2 根据 n 只麻雀的初始位置计算相应的适应度值，并对它们排序。

步骤 3 更新发现者位置：

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^t \cdot \exp\left(-\dfrac{i}{\partial \cdot iterm_{\max}}\right), & R_2 < ST \\ X_{i,j}^t + Q \cdot L, & R_2 \geq ST \end{cases} \tag{2-30}$$

上式中，$X_{i,j}^t$ 表示在第 t 次迭代时第 i 只麻雀在第 j 维中的位置信息，这里 $j=1,2,\ldots,d$。$\partial \in (0,1]$ 和 $R_2 \in [0,1]$ 都是均匀分布的随机数，R_2 和 $ST\left(ST \in [0.5,1]\right)$ 分别表示预警值和安全值。Q 是服从标准正态分布的随机数。L 表示元素均为 1 的一个 $1\times d$ 的矩阵。

步骤 4 更新加入者位置：

$$X_{i,j}^{t+1}=\begin{cases} Q\cdot\exp\left(\dfrac{X_{worst}^{t}-X_{i,j}^{t}}{i^{2}}\right), i>n/2 \\ X_{P}^{t+1}+\left|X_{i,j}^{t}-X_{P}^{t+1}\right|\cdot A^{+}\cdot L \text{ ，其他} \end{cases} \tag{2-31}$$

其中， X_{p}^{t+1} 是 $t+1$ 时刻发现者的最佳位置， X_{worst}^{t} 则表示当前全局最差的位置。A 表示 $1\times d$ 的矩阵，其中每个元素随机赋值为 1 或-1，并且 $A^{+}=A^{T}\left(AA^{T}\right)^{-1}$。

步骤 5 更新警戒者位置：

$$X_{i,j}^{t+1}=\begin{cases} X_{best}^{t}+\beta\cdot\left|X_{i,j}^{t}-X_{best}^{t}\right| , f_i>f_g \\ X_{i,j}^{t}+K\cdot\left(\dfrac{\left|X_{i,j}^{t}-X_{best}^{t}\right|}{\left(f_i-f_w\right)+\varepsilon}\right) , f_i=f_g \end{cases} \tag{2-32}$$

其中 X_{best}^{t} 是当前的全局最优位置。β 作为步长控制参数，是服从标准正态分布的随机数。$K\in[-1,1]$ 是一个均匀分布的随机数， f_i 则是当前麻雀个体的适应度值。 f_g 和 f_w 分别是当前全局最佳和最差的适应度值。 ε 为常数，以避免分母出现零。

步骤 6 根据麻雀的更新的位置重新计算每个麻雀个体的适应度值。

步骤 7 根据适应度值及最大迭代次数判断是否满足停止条件，满足则退出，输出结果。否则，重复执行步骤 2-6。

2.5 变量筛选与合成

为了将与机场旅客吞吐量相关联的关键词的百度搜索量合成为一个综合指数，本文进行关键词的筛选和搜索信息加权整合。首先利用主观选词法在百度指数网站(https://index.baidu.com)初选机场自身及机场所在地旅游相关的核心搜索关键词后，再利用范围选词法通过百度指数的需求图谱功能对关键词进行拓词，并获取相应关键词的日度搜索量。为统一搜索关键词序列与机场旅客吞吐量的统计口径，将日度搜索量进行加总，整合为月度数据。然后利用平均影响值(Mean Impact Value, MIV)[47]方法对关键词进行初步筛选，再利用时差相关分析法进一步优选，并将关键词信息与原始数据的相关系数作为权重，将筛选得到的各关键词搜索量加权合成为综合搜索指数。

2.5.1 平均影响值(MIV)

MIV 是判断输入变量对输出变量影响程度大小的指标，通过同比例增减输入变量的值，得到不同的输出变量，将两次输出变量的差值定义为影响值，则差值的均值即为 MIV。其具体计算过程如下：

将训练样本集中变量 X_i 的取值按 10%增加或减少，得到两个新样本 A_1 和 A_2，带入 BP 神经网络训练得到两组预测值，计算该两组预测值的差值 IV_i，求其平均值 MIV_i 即为平均影响值；

按如下公式求得自变量 X_i 的相对贡献率：

$$v_i = \frac{|MIV_i|}{\sum_{i=1}^{n}|MIV_i|} \tag{2-33}$$

本文中 v_i 表示第 i 个关键词对机场旅客吞吐量的相对贡献率，n 表示关键词个数。

2.5.2 时差相关分析(TDR)

时差相关分析法是测量时间序列之间的领先、同步或滞后关系的常用方法。设 l 为两序列 $\{x_t\}$ 与 $\{y_t\}$ 的时间差，$\bar{x}$ 与 $\bar{y}$ 分别为两序列的平均值。$\{x_t\}$ 移位 l 期后与 $\{y_t\}$ 的相关系数 r_l 可以通过以下公式计算：

$$r_l = \frac{\sum_{t=1}^{m}(x_{t+l}-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_{t=1}^{m}(x_{t+l}-\bar{x})^2(y_t-\bar{y})^2}},\ l=0,\pm1,\pm2,...,\pm p \tag{2-34}$$

其中，m 为样本个数，当 l=0 时，r_0 表示 x_t 与 y_t 的同阶相关系数，即皮尔逊相关系数。当 $l<0$ 时，表示 x_t 滞后 l 期与 y_t 的相关系数。本文将第 i 个关键词的搜索量 s_i 与机场旅客吞吐量 y 做相关性分析，根据 s_i 取不同滞后期后与序列 y 的时差相关系数大小，可确定出最优的滞后期数 k_i。

2.5.3 合成综合搜索指数

筛选出关键词后需要将所有关键词的百度搜索指数合并为一个综合指数，将每个关键词在最佳滞后期数内的搜索量累积相加后与机场旅客吞吐量的皮尔逊相关系数作为权重，对关键词搜索信息进行加权整合，从而形成与机场旅客吞吐

量相关的综合搜索指数。合成综合搜索指数序列S_t的具体公式如下：

$$S_t = \sum_{i=1}^{d} \rho_i \sum_{j=1}^{k_i} s_i^{t-j} , \tag{2-35}$$

其中，d为筛选出的关键词个数，s_i^{t-j}为第i个关键词滞后j期的搜索信息，ρ_i为第i个关键词在最佳滞后期内的搜索量总和与机场旅客吞吐量的皮尔逊相关系数。

3 基于分解集成的我国港口集装箱吞吐量预测

本章在分解集成预测框架下，提出了一种新的组合预测模型，即用 EEMD 算法将原始序列分解为若干子序列后，将不同频率的序列用 PSO-LSSVR 来预测，而对残差趋势项用多项式回归(Polynomial Regression,PR)来拟合，接着用感知机模型对不同频率序列的预测值进行非线性集成，与对应残差项的预测值进行加总从而得到最终预测结果。

3.1 分解-集成组合预测模型框架

本章中所用的基于”分解-集成”框架下的组合预测模型，主要包括分解、各模态预测和综合集成三个部分。其中，分解的主要目的是简化预测的任务，将其划分成易于预测的各个子任务；集成主要是为了形成原始数据的最终预测结果，而且集成学习对最终预测的结果非常重要。本文使用的分解-集成框架下的预测模型，该方法相比传统的预测方法，可以获得比较好的性能。在这里，我们首先使用 EEMD 算法将原始集装箱吞吐量数据进行分解；其次，将带有趋势项的残差序列进行传统多项式回归预测，对其余模态进行 PSO-LSSVR 预测；最后，对各个预测结果进行非线性集成。图 3.1 为该方法的总体框架预测流程图。

本章中在使用 EEMD 将白噪声标准差确定为 0.2，添加白噪声的次数确定为 100，根据数据自身特性，分解为若干个子列，逐一进行分析与预测。

具体的预测过程包括以下三个步骤：

步骤 1 分解。使用 EEMD 算法将原始数据分解为 N 个相对简单且有意义的本征模态函数 IMF_1、IMF_2、IMF_N 和残差序列 Res。

步骤 2 模态预测。不同频率的 IMF 用 PSO-LSSVR 进行预测，具体根据原始数据的特征确定相应的滞后期数，将滞后的期数作为 PSO-LSSVR 的输入，后一期作为输出，根据滚动预测的原则得出相应测试集 IMF 的预测值。对带有趋势的残差项用 PR 进行预测，根据训练集确定 PR 的回归系数，得到的的一元多项式用来预测残差项。

步骤 3 集成预测。将 PSO-LSSVR 预测的各子模态的结果用感知机模型(PM)进行非线性集成，将集成预测的结果与对应残差项的预测值相加，形成最终的预测结果。

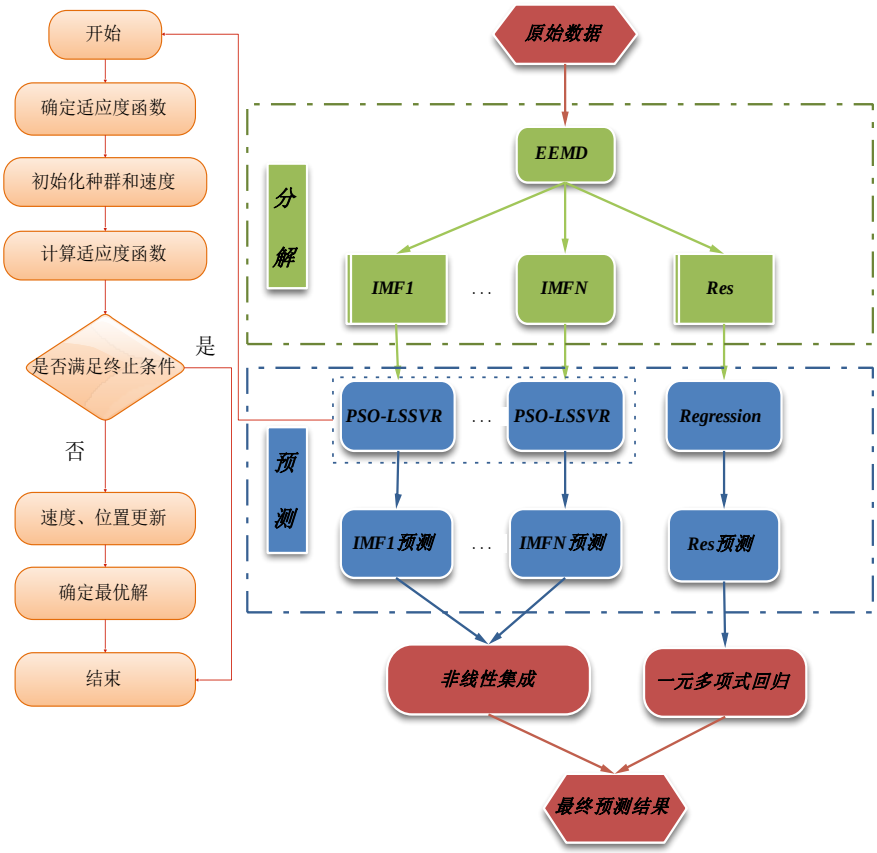

图3.1 分解-集成的总体框架预测流程

3.2 实证分析

3.2.1 数据与评价准则

为使所选的数据具有代表性，本文以 2019 年国家交通运输部发布的全国港口集装箱吞吐量数据前十的十大港口，即广州、青岛、大连、天津、厦门、苏州、营口、上海、宁波-舟山以及深圳的港口集装箱吞吐量数据(数据来源于 Wind 数据库及中华人民共和国交通运输部)为研究对象，其中广州港、青岛港、大连港、天津港、厦门港、上海港、深圳港的数据涵盖了 2001 年 1 月至 2019 年 12 月期间共计 228 个观测值，宁波-舟山港的数据涵盖了 2006 年 1 月至 2019 年 12 月共计 168 个观测值，苏州港的数据涵盖了 2011 年 1 月至 2019 年 12 月共计 108 个观测值，营口港的数据涵盖 2009 年 1 月至 2019 年 12 月共计 132 个观测值。

图 3.2(a)为十大港口集装箱吞吐量的变化曲线，各个数据的趋势变动、波动幅度都略有不同。为了捕捉不同数据的差异性以及找出有显著性特征差异的代表性港口，本文运用 K-Means 聚类将数据进行分类，每一类中选出最具有代表性的港口数据进行模型检验，用不同特征的港口数据来检验本文所建模型的稳健

性。在进行 K-Means 聚类之前，为克服不同港口集装箱数据长度不同的问题，我们用零值将具有较少样本点的宁波-舟山港、苏州港、营口港的数据补齐至与广州港、上海港的样本数相同，再通过计算并比较各港口集装箱吞吐量数据间欧式距离的大小，将距离较近的数据归为一类，可将数据分为 3 类，选出每一类中距离类中心最近的港口数据作为代表来进行预测。

图 3.2(b)分别给出了三类中最具有代表性的广州、营口、上海港口集装箱吞吐量的时序图，观察发现广州、营口、上海的特征差异明显，分别代表了不同频率及不同波动大小的三类数据。将广州港、营口港、上海港的数据集分为训练集和测试集，上海港和广州港 2001 年 1 月至 2018 年 12 月共 216 个观测值、营口港 2009 年 1 月至 2018 年 12 月共 120 个观测值分别作为训练集，此数据用来模型训练。以上海港、广州港和营口港 2019 年 1 月至 2019 年 12 月共 12 个数据作为测试集。本文根据原始序列的自相关程度大小选取了滞后 6 期的吞吐量数据作为输入预测下一月的数据。

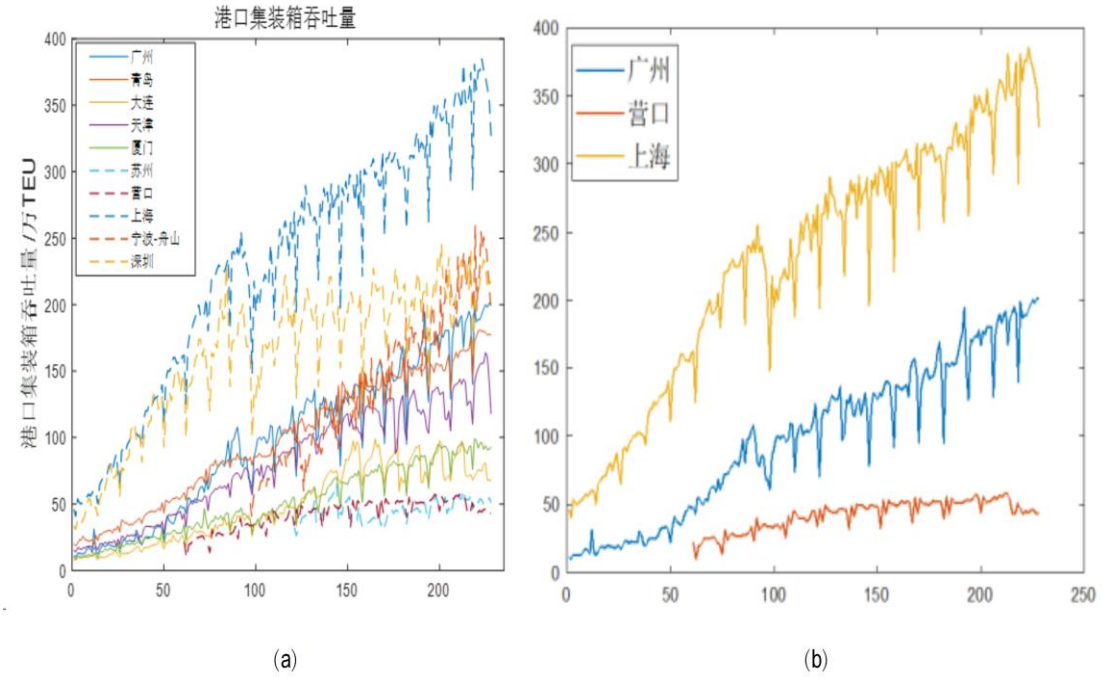

图3.2 港口集装箱吞吐量

通常有多种误差测量标准可用于评估模型的预测性能。但是，相关研究表明，没有通用的标准公式可以评估预测方法的有效性。本文采用四种主要的评估标准来评估预测性能：平均绝对百分比误差(MAPE)，均方根误差(RMSE)，平均绝对误差(MAE)和方向统计量 D_{stat} 。较低的 MAPE、RMSE、MAE 表示较好的水平预

测性能，较高的 D_{stat} 表示较好的方向预测性能。MAPE、RMSE、MAE、D_{stat} 的计算公式分别如下：

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{3-1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3-2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{3-3}$$

$$D_{stat} = \frac{1}{n}\sum_{i=1}^{n}A_i \times 100\% \tag{3-4}$$

其中，$\hat{y}_i$ 表示预测值，y_i 表示实际值，n 表示观测样本的数目。式(3-4)中，当 $(\hat{y}_{i+1} - y_i)(y_{i+1} - y_i) \geq 0$ 时，A_i 的返回值为 1，否则为 0。

根据表 3.1 的 ADF 检验结果所示，在 1%、2%和 5%的显著性水平下，上海、广州、营口的集装箱吞吐量数据的 p 值大于 0.05，故接受单位根原假设，认为原始数据是不平稳的。因此，本文研究的集装箱吞吐量数据具有典型的非平稳性特征。

表 3.1 数据的 ADF 单位根检验

	上海港		广州港		营口港	
	t 统计量	P 值	t 统计量	P 值	t 统计量	P 值
ADF 检验统计量	-1.9723	0.299	0.3142	0.9786	-2.5816	0.0996
1% level	-3.4610		-3.4607		-3.4861	
5% level	-2.8749		-2.8748		-2.8859	
10% level	-2.5740		-2.5739		-2.5798	

3.2.2 模型结果分析

在本文的分析中，为了比较各个模型的预测性能与精度，提出了 ES、LSSVR、PSO-LSSVR 作为单一的基准模型，而 EMD-LSSVR-PM、EMD-PSO-LSSVR-PM、EMD-PSO-LSSVR/PR-PM、EEMD-LSSVR-PM、EEMD-PSO-LSSVR-PM 作为分解-集成框架下的基准模型来预测上海、广州、营口港的数据，与本文提出的 EEMD-PSO-LSSVR/PR-PM 模型进行比较。

(1)首先，用 EEMD 将上海、广州、营口港口集装箱吞吐量数据进行分解，将上海、广州、营口港口集装箱吞吐量数据分别分解为 6 个本征模态分量与 1 个

残差趋势分量，如图 3.3、3.4、3.5 所示。

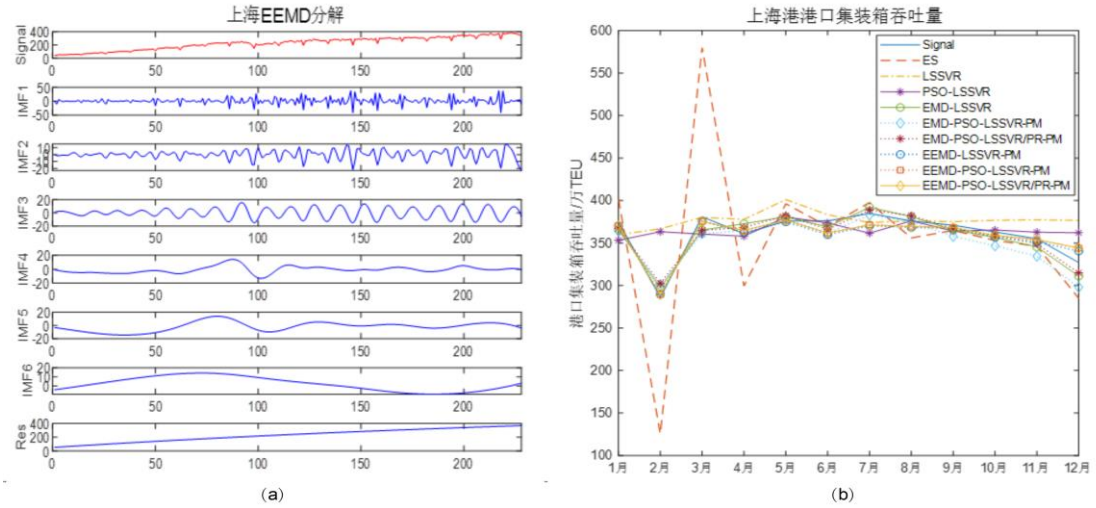

图3.3 上海港港口集装箱吞吐量EEMD分解与模型预测结果对比图

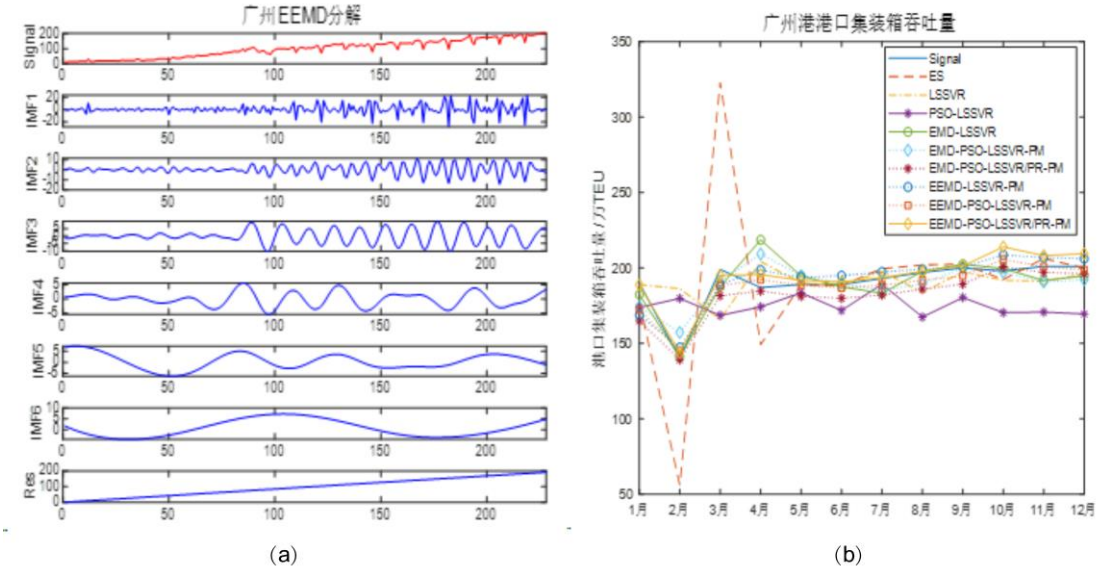

图3.4 广州港口集装箱吞吐量EEMD算法分解与模型预测结果对比图

(2)接着，我们分别利用 PSO-LSSVR 和 PR 对 EEMD 算法分解后的 6 个 IMFs 分量和 1 个残差分量进行预测。

(3)然后，将各 IMF 分量和残差分量的预测结果进行集成汇总。集成汇总的过程采用 PM 模型集成，然后用简单相加的集成方式得到最终的预测结果，即将各模态的预测分量用 PM 进行非线性集成，具体将 6 个 IMFs 预测得到的值作为 PM 的输入，而用原始数据减去 EEMD 算法分解所得趋势残差项 Res 的值作为输出来训练模型，如此将各 IMF 预测的数据进行非线性集成，得到 6 个 IMFs 的非线性集成预测值。然后再将每期各 IMF 的总预测值与对应残差分量的预测值进行加总，得到最终的预测结果。图 3.3(b)、图 3.4(b)、图 3.5(b)分别展示了模型对

上海、广州、营口 2019 年 1 月至 2019 年 12 月集装箱吞吐量数据预测结果的对比图。

(4)为了验证本文提到的分解-集成预测方法的有效性，本文用一个单一模型 ES 、 LSSVR 、 PSO-LSSVR 与数个组合模型 EMD-LSSVR-PM 、 EMD-PSO-LSSVR-PM 、 EMD-PSO-LSSVR/PR-PM 、 EEMD-LSSVR-PM 、 EEMD-PSO-LSSVR-PM，与本文提出的模型进行比较，利用 MAPE、RMSE、MAE、D_{stat}，将模型预测结果计算出的误差指标进行对比，具体如表 3.2 所示。

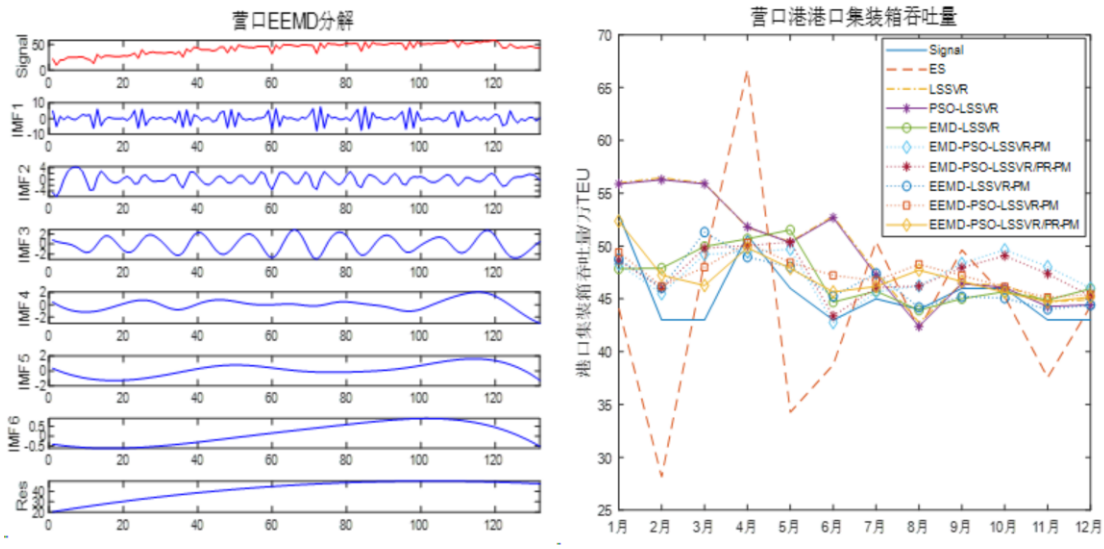

图3.5 营口港港口集装箱吞吐量EEMD分解与模型预测结果对比图

表 3.2 不同模型的预测效果比较

港口	模型	RMSE	MAPE(%)	MAE	D_{stat} (%)
上海	ES	77.8783	13.9893	47.8127	91.67
	LSSVR	31.2546	6.6216	21.9072	50
	PSO-LSSVR	28.3458	5.5851	18.4713	66.67
	EMD-LSSVR-PM	9.8539	2.5828	9.1333	91.67
	EMD-PSO-LSSVR-PM	10.7000	2.3752	8.4982	91.67
	EMD-PSO-LSSVR/PR-PM	9.1206	2.3726	8.5758	83.33
	EEMD-LSSVR-PM	9.0531	3.9421	7.4124	91.67
	EEMD-PSO-LSSVR-PM	8.4219	1.8826	6.8006	91.67
	EEMD-PSO-LSSVR/PR-PM	8.2918	1.7984	6.4561	83.33
广州	ES	44.8494	13.5548	23.7557	91.67
	LSSVR	17.9588	6.9775	12.1086	66.67
	PSO-LSSVR	19.4071	8.213	14.6266	66.67
	EMD-LSSVR	11.0059	3.8715	7.4234	91.67

续表 3.2 不同模型的预测效果比较

	EMD-PSO-LSSVR-PM	10.9934	4.8538	8.8286	91.67
	EMD-PSO-LSSVR/PR-PM	10.7588	4.7578	8.6625	58.33
广州	EEMD-LSSVR-PM	9.0857	3.9638	7.4331	91.67
	EEMD-PSO-LSSVR-PM	7.219	2.9298	5.5302	91.67
	EEMD-PSO-LSSVR/PR-PM	6.4241	2.4602	4.6924	83.33
	ES	8.2183	14.3292	6.6179	91.67
	LSSVR	6.8166	10.5676	4.6617	75.00
	PSO-LSSVR	6.2953	9.6577	4.2464	75.00
	EMD-LSSVR	3.4927	5.8114	2.624	83.33
营口	EMD-PSO-LSSVR-PM	3.4385	6.5939	2.9891	91.67
	EMD-PSO-LSSVR/PR-PM	3.391	6.4215	2.9014	75.00
	EEMD-LSSVR-PM	3.164	5.2712	2.3894	91.67
	EEMD-PSO-LSSVR-PM	2.9664	5.7871	2.5892	75.00
	EEMD-PSO-LSSVR/PR-PM	2.3204	4.4666	1.973	83.33

根据表 3.2 中各模型的预测效果来看，我们可以得到以下结论：

(1)与单一传统预测方法 ES 相比，不论是上海港、广州港还是营口港的预测结果，智能预测方法、基于分解-集成框架下的预测方法都具有更小的 MAPE、RMSE、MAE，即这些模型的预测精度更高。说明对于非平稳数据，智能预测方法、分解-集成框架下的预测方法更具可行性。

(2)比较单一智能预测方法 LSSVR 与通过启发式算法对参数进行优化后的智能预测方法 PSO-LSSVR，模型的预测精度得到了不同幅度的提高(除广州港)，即 PSO 对 LSSVR 模型中惩罚参数 C 与核参数 σ^2 优化选择的结果较为显著，进一步说明启发式算法在模型中参数选择的有效性。

(3)从 EMD-LSSVR、EMD-PSO-LSSVR-PM、EMD-PSO-LSSVR/PR-PM 和 EEMD-LSSVR-PM、EEMD-PSO-LSSVR-PM、EEMD-PSO-LSSVR/PR-PM 不同分解-集成的组合模型预测结果来看，首先，由于 EEMD 算法克服了 EMD 算法分解过程中模态混叠的问题，故基于 EEMD 算法下组合模型得到了更高的预测精度。其次，对 EEMD 算法分解后得到的不同频率的六个 IMFs 与残差项 Res，各 IMF 采用 PSO-LSSVR 进行预测，带有趋势的残差项采用多项式回归进行拟合。这是因为，趋势残差项的非线性程度不高，波动性也不高，采用传统的趋势预测方法反而比智能预测方法更优。事实证明，针对不同的数据特征，应选取相适应的预测方法，”分而治之”策略下能进一步提高模型的预测精度。最后，对于不同

频率子序列的预测值采用 PM 进行非线性集成，最终得到 EMD-PSO-LSSVR/PR-PM 为最优模型。

4 基于二次分解集成框架下的航空客流预测

本章在一次分解集成预测框架的前提下，提出 EEMD-VMD-SE-CS-BP 二次分解模型，二次分解不仅能进一步捕捉到时间序列信息，同时能够降低子序列预测的难度。本章主要包括：4.1 节介绍了二次分解的模型框架；4.2 节详细介绍了数据来源及其评测标准；4.3 节介绍了 EEMD-VMD-SE-CS-BP 方法的实现过程及模型的结果分析。

4.1 模型框架

基于“分解-集成”的思想，作为对一次分解的补充，提出了一种二次分解的组合预测新模型。为了解决由非平稳，非线性得航空客流数据复杂性问题，利用分解技术将原始序列分解成不同得子序列，根据数据得复杂性特征，将具有相同特征属性分量进行整合重构，一次分解方法的不能充分提取数据的高频特性。文章提出了一种新的二次分解技术，进一步分解重构产生的高频、中频分量，以降低了高、中的复杂度，提高每个序列的预测精度。

首先分解原始机场旅客吞吐量序列，为减小预测误差累积，将分解得到的子序列重构为高、中、低频序列。再对高、中频序列再次分解为若干子序列，降低序列复杂度，然后对各子序列分别进行预测，高频、中频序列的预测值分别进行集成后，对高、中、低频序列的预测值进行非线性集成，得到最终预测值。图 4.1 为本文的总体预测框架流程图。具体的预测过程主要包括如下五个部分：

(1) 数据的分解与重构

步骤 1 数据的分解。利用 EEMD 分解方法对将机场旅客吞吐量数据进行分解，得到旅客吞吐量数据的 m 个本征模态函数 IMF_1、…、IMF_m 及残差 Res。

步骤 2 子序列重构。分别计算 $m+1$ 个子序列的样本熵值，根据样本熵值，如果 $SE>1$，子序列合并为高频序列 H，$0.5<SE<1$ 的子序列合并为中频序列 M，$SE<0.5$ 合并为低频序列 L。

(2) 二次分解

步骤 3 为更进一步捕捉序列的波动、变化趋势，将高频 H、中频 M 序列采用 VMD 算法进行二次分解，由于 VMD 算法的分解个数必须预先设定，为防止序列过分解，我们计算子序列的中心频率变化得到高频 H、中频 M 序列最佳的

分解个数 p 、 q 。

(3)数据的预测

步骤 4 将高、中频分解得到的 p 个、q 个子序列以及低频序列均采用 CS-BP 模型进行预测，得到高频序列 p 个子序列的预测值、中频序列 q 个子序列的预测值及低频序列的预测值。

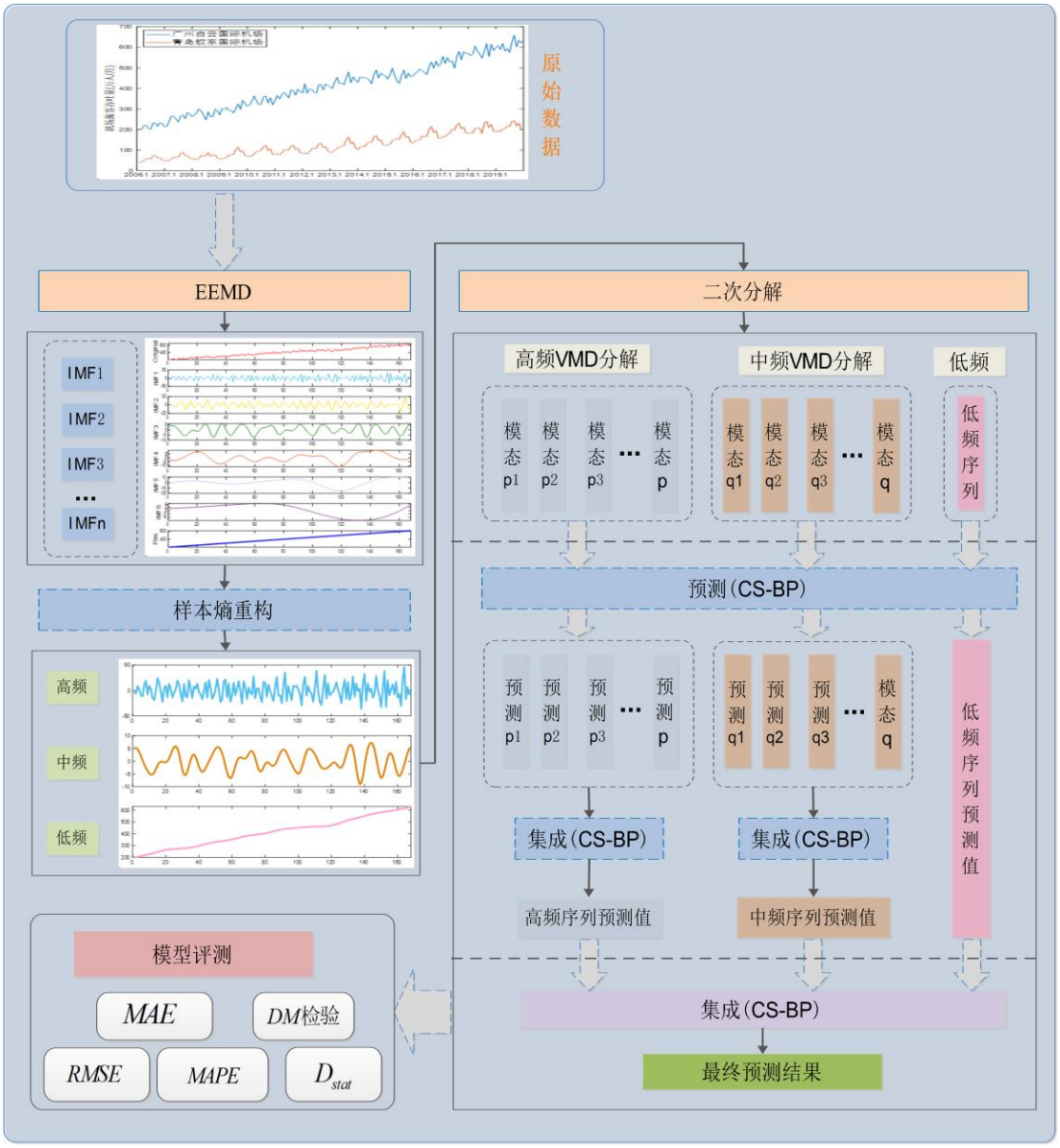

图4.1 二次分解预测模型流程图

(4)数据集成

步骤 5 将高频 p 个子序列的预测值和中频 q 个子序列的预测值分别进行集成。高频序列经过 VMD 分解得到的 p 个子序列的预测值作为神经网络的输入，高频序列作为输出，建立 CS-BP 模型进行集成，中频序列同理。

步骤 6 最终非线性集成。将训练集样本的高、中、低频序列预测值作为输入，观测值作为输出，训练 CS-BP 模型，对将测试集的值带入模型中进行集成，即得到了最终的预测值。

(5)模型比较

步骤 7 通过建立基准模型与文章提出的二次分解集成模型进行水平精度、方向精度的比较，且通过统计检验来检验模型的优劣。

4.2 数据来源及评测标准

本文选取 2006 年 1 月至 2019 年 12 月的广州白云国际机场、青岛胶国际机场的月度客运量为研究对象，数据来源于 Wind 数据库。数据共 168 个观测值，将 2006 年 1 月至 2016-12 月共 132 个观测值作为样本内数据，用来训练模型，将 2017 年 1 月至 2019 年 12 月共 36 个观测值作为样本外数据，用来检验模型。

为了更全面、客观地评估模型的预测性能，本文采用以下三种水平指标：均方根误差(RMSE)、平均绝对百分比误差(MAPE)、平均绝对误差(MAE)，以及方向预指标： D_{stat} 。具体公式见 3.2.1 中介绍。

航空客运数据呈现出非线性、非平稳、等复杂特征，图 4.1 是关于广州白云机场与青岛胶东机场客流数据的变化波动图，表 4.1-4.2 是该数据相关统计量，广州和青岛机场的数据未通过 ADF 单位根检验，证实机场数据为非平稳的时间序列数据。

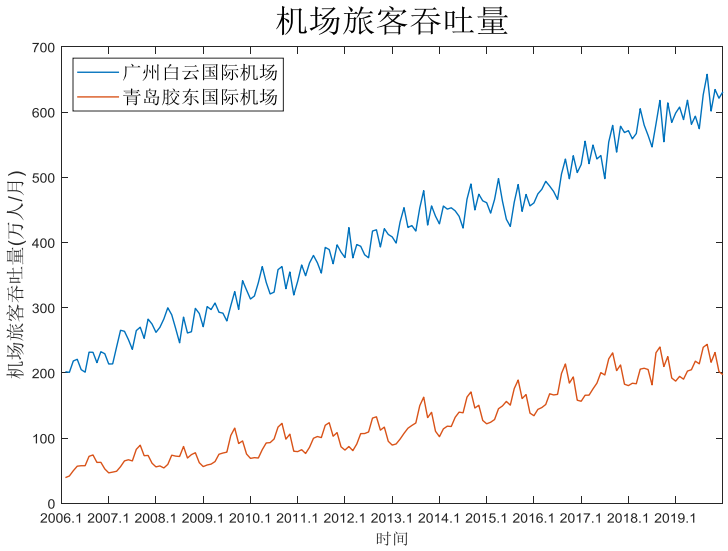

图4.2 机场旅客吞吐量

表 4.1 描述性统计量

序列 \ 统计量	均值	最大值	最小值	中位数	标准差	峰度	偏度
广州白云国际机场	411.8	658.1	200.9	420.5	120.6	1.95	0.07
青岛胶东国际机场	125.5	243.9	39.5	116.2	54.5	1.96	0.40

表 4.2 ADF 单位根检验——平稳性检验

	广州白云国际机场		青岛胶东国际机场	
	t 统计量	p 值	t 统计量	p 值
ADF 检验统计量	-0.0094	0.9556	1.5294	0.9993
1%level	-3.4728		-3.4728	
5%level	-2.8801		-2.8801	
10%level	-2.5767		-2.5767	

4.3 实证结果与分析

4.3.1 二次分解模型建立

(1)首先，用 EEMD 将广州机场、青岛机场旅客吞吐量数据进行分解，得到原始序列分解为 6 个本征模态函数与 1 个残差趋势分量，如图 4.3-4.4 所示。

(2)其次，我们计算每个序列的样本熵值。根据样本熵值，我们将 EEMD 分解得到的子序列划分为高频、中频、低频。根据样本熵的特点，值越大表示序列越复杂。其中高频信号为样本熵值大于 1 的序列，中频信号为样本熵值大于 0.5 的序列，低频信号则为样本熵值小于 0.5 的序列。如图 4.5 和图 4.6 所示。

(3)对高、中频的序列进行二次分解，进一步提取数据的变化特征，由于 VMD 分解层数需要人为确定，有一种较为便捷的模态数 K 的选取方法，其基本思想是考虑 VMD 分解之后各模态的中心频率互不相同，如果出现相邻两个模态之间中心频率相近，则认为出现过分解，当前分解层数减 1 作为最佳分解层数。如表 4.2、表 4.3 所示，高频、低频序列分别分解为 8 个子模态和 6 个子模态时，最小的中心频率间距明显发生变化，表示数据已经过分解了，所以高、中频序列应分解为 7 个和 3 个子序列，如图 4.7、图 4.8 所示。对高频、中频进行 VMD 分解后，分别计算各子序列的样本熵值，如表 4.4、表 4.7 所示，高频、中频序列均被分解为样本熵值较低，即复杂度较低的子序列。

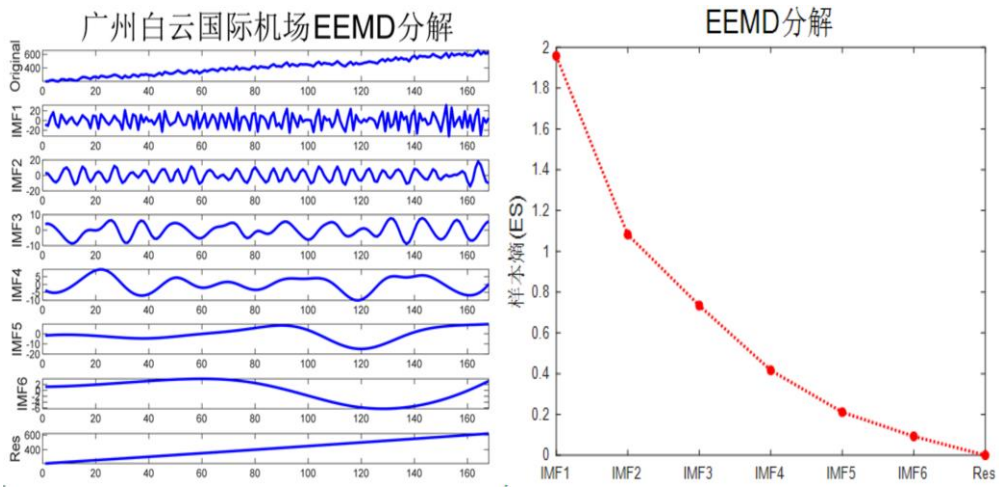

图4.3 广州白云国际机场EEMD分解与样本熵值

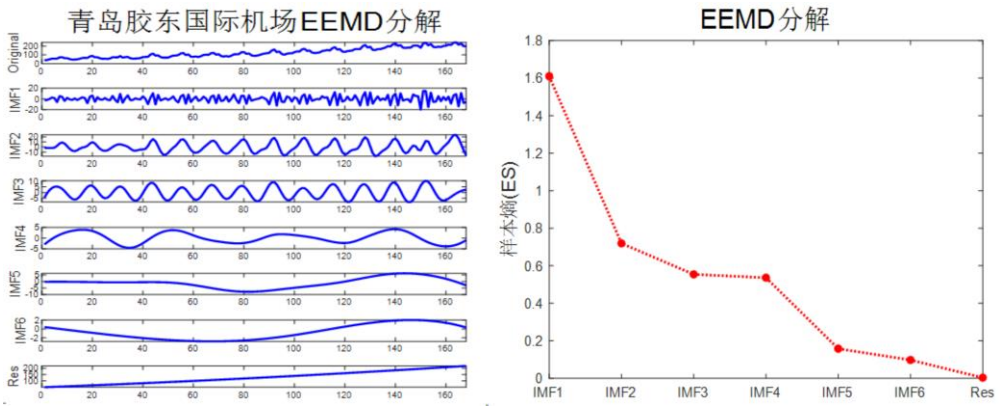

图4.4 青岛胶东国际机场EEMD分解与样本熵

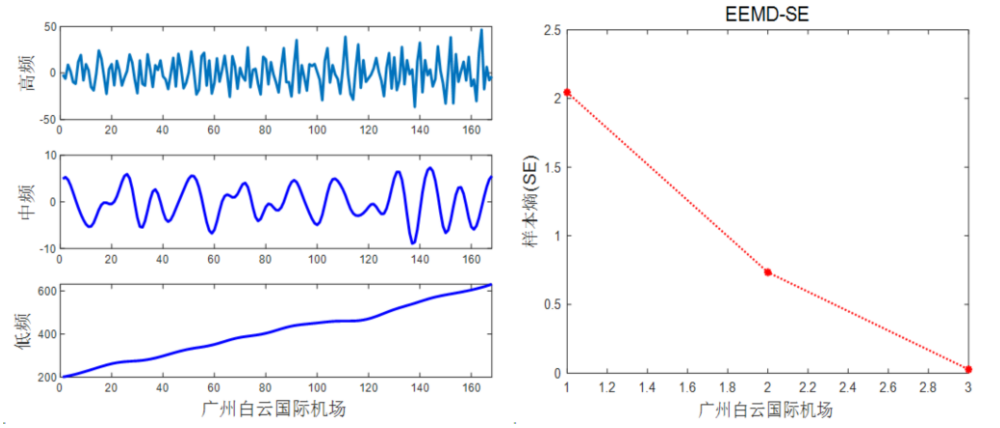

图4.5 广州白云国际机场EEMD分解与样本熵值

(4)对每个子序列用 CS-BP 方法进行预测，为了判断每个子序列的滞后期数，文章采用试错法，使数据自适应的选择最佳滞后期数。分别对每个序列滞后 2-10 期的预测值进行评估，本文采用 RMSE 对不同滞后期数的预测值进行评估。如广州白云国际机场，高频各子序列最佳滞后为 3、3、4、8、4、6、3，中频各子序列最佳滞后期为 8、8、4，低频序列最佳滞后期为 3。如表 4.8-4.9 所示(此表

所列数据均为最佳滞后期预测精度)，高频序列经过 VMD 得到的子序列在水平和方向精度上均有所提高。

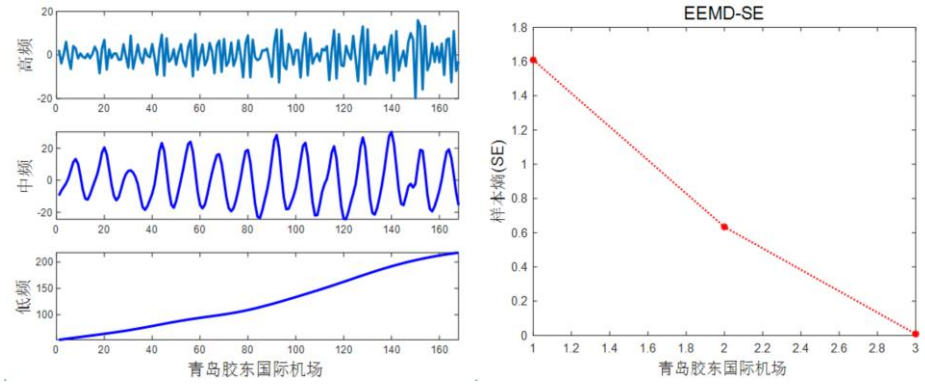

图4.6 青岛胶东国际机场EEMD分解与样本熵值

(5)将上述(4)中得到的高频序列 VMD 分解得到 7 各子序列的预测值作为输入变量，高频序列作为输出变量，建立 CS-BP 模型，对子序列进行非线性集成，得到高频序列的预测值。中频序列同样将 3 个子序列的预测值进行非线性集成，，得到中频序列的预测值。

(6)将(5)中通过集成得到的高频、中频序列预测值与(4)中得到的低频序列的预测值进行非线性集成，高、中、低频序列的预测值作为 CS-BP 模型的输入变量，训练集的观测值作为输出变量，训练模型，对测试集数据进行预测。

表 4.2 广州白云国际机场高频序列 VMD 分解中心频率-高频序列

K1	中心频率/HZ								最小间距
2	0.1488	0.2519							0.1031
3	0.1166	0.1705	0.3314						0.0539
4	0.1113	0.1664	0.2515	0.4121					0.0551
5	0.1604	0.1206	0.2380	0.3288	0.4162				0.0398
6	0.1043	0.1347	0.1688	0.2495	0.3312	0.4167			0.0304
7	0.1048	0.1349	0.1685	0.2469	0.2818	0.3326	0.4168		0.0301
8	0.1037	0.1322	0.1674	0.1985	0.2501	0.3310	0.4165	0.4169	0.0004

表 4.3 广州白云国际机场中频序列 VMD 分解中心频率-中频序列

K1	中心频率/HZ				最小间距
2	0.047	0.0888			0.0418
3	0.0475	0.0938	0.0812		0.0125
4	0.0473	0.0831	0.077	0.1067	0.0062

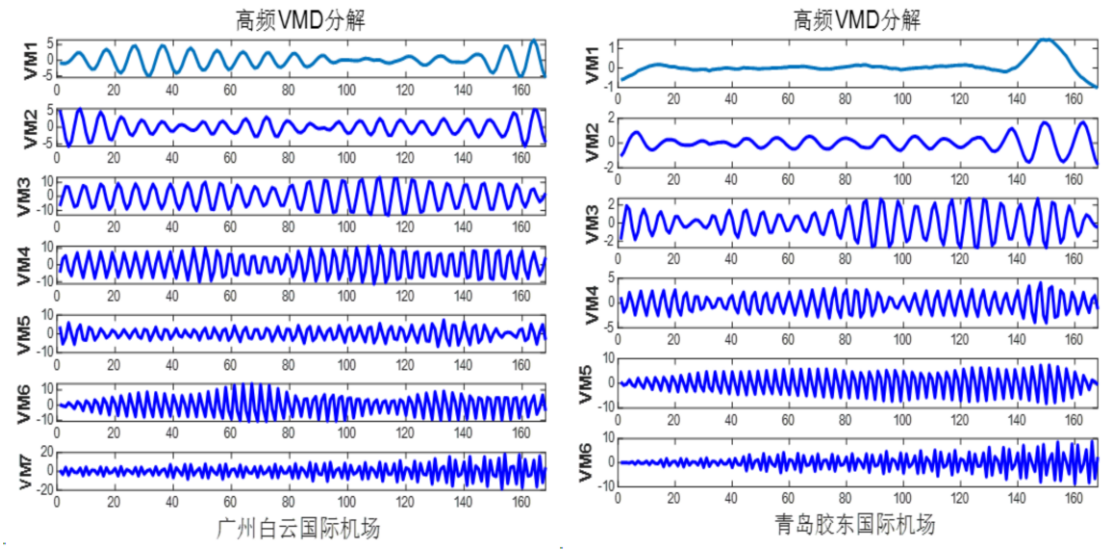

图4.7 广州、青岛高频序列VMD分解子序列

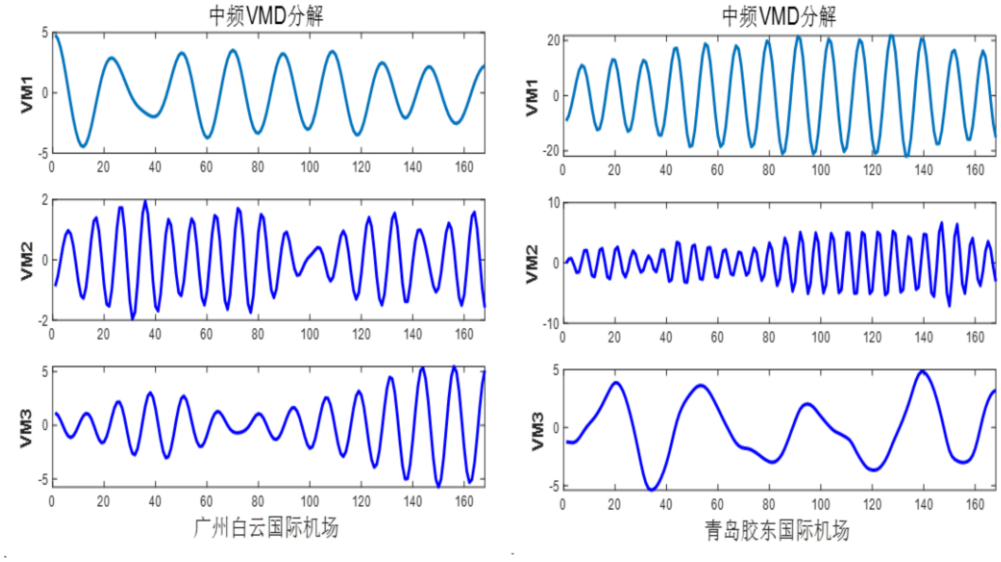

图4.8 广州、青岛高频序列VMD分解子序列

表 4.4 广州白云国际机场高频、中频序列 VMD 分解样本熵值比较

	高频序列	高频序列 VMD 分解		中频序列	中频序列 VMD 分解	
		最大值	最小值		最大值	最小值
样本熵	2.04	0.59	0.22	0.73	0.60	0.53

表 4.5 青岛胶东国际机场高频序列 VMD 分解中心频率-高频序列

K1	中心频率/HZ				最小间距
2	0.1309	0.3186			0.1877
3	0.0647	0.1738	0.3337		0.1091
4	0.0631	0.1671	0.2575	0.3409	0.0834

续表 4.5 青岛胶东国际机场高频序列 VMD 分解中心频率-高频序列

5	0.0565	0.1614	0.2468	0.3310	0.4155			0.0842
6	0.0231	0.0846	0.1696	0.2553	0.3331	0.4164		0.0615
7	0.0239	0.0830	0.1689	0.2536	0.3293	0.3379	0.4165	0.0085

表 4.6 青岛胶东国际机场中频序列 VMD 分解中心频率-中频序列

K1	中心频率/HZ				最小间距
2	0.0735	0.1694			0.0959
3	0.0830	0.1645	0.0769		0.0815
4	0.0748	0.0837	0.1656	0.1122	0.0089

图 4.7 青岛胶东国际机场高频、中频序列 VMD 分解样本熵值比较

	高频序列	高频序列 VMD 分解		中频序列	中频序列 VMD 分解	
		最大值	最小值		最大值	最小值
样本熵	1.61	0.64	0.21	0.63	0.52	0.26

表 4.8 广州白云国际机场高频序列预测值及 VMD 分解预测值比较

	高频序列	高频序列 VMD 分解						
		VM1	VM2	VM3	VM4	VM5	VM6	VM7
RMSE	14.69	0.25	0.31	0.27	0.30	0.34	0.29	0.42
MAPE(%)	115.10	25.41	25.06	16.56	11.23	22.60	6.38	3.31
MAE	12.46	0.20	0.24	0.22	0.23	0.27	0.24	0.32
D_{stat} (%)	80.56	88.89	97.22	97.23	91.67	97.22	97.22	97.22

表 4.9 青岛胶东国际机场高频序列预测值及 VMD 分解预测值比较

	高频序列	高频序列 VMD 分解					
		VM1	VM2	VM3	VM4	VM5	VM6
RMSE	5.53	0.01	0.15	0.11	0.17	0.14	0.15
MAPE(%)	82.83	0.55	16.48	12.37	9.93	9.36	3.45
MAE	4.05	0.01	0.12	0.09	0.13	0.12	0.12
D_{stat} (%)	91.67	94.44	94.44	94.44	97.22	97.22	97.22

4.3.2 结果分析

本文将提出的模型为 EEMD-SE-VMD- -CS-BP，文章建立四类基准模型进行比较，传统时间序列预测模型：整合移动平均自回归模型(ARIMA)；单一神经网络预测模型：BP 神经网络预测模型、布谷鸟搜索算法优化 BP 神经网络模型

CS-BP；混合预测模型：一次分解-集成预测模型 EEMD-BP，一次分解-重构预测模型 EEMD-SE-BP、EEMD-SE-CS-BP，二次分解-重构预测模型 EEMD-VMD-SE-BP。

表 4.10 广州白云国际机场模型预测结果比较(1 步预测)

模型	RMSE	MAPE(%)	MAE	D_{stat} (%)
ARIMA	32.5943	4.7992	27.1859	63.89
BP	29.1222	4.0422	23.7210	58.33
CS-BP	27.2285	3.6459	21.2800	77.78
EEMD-BP	19.2779	2.5934	15.2927	80.56
EEMD-SE-BP	16.4686	2.2854	13.4745	80.56
EEMD-SE-CS-BP	14.7004	2.2421	13.0219	80.56
EEMD-VMD-SE-BP	8.2034	1.1922	6.9052	94.44
EEMD-VMD-SE-CS-BP	4.7728	0.6781	3.9434	97.22

表 4.11 广州白云国际机场模型预测结果比较(3 步预测)

模型	RMSE	MAPE(%)	MAE	D_{stat} (%)
ARIMA	32.5943	4.7992	27.1859	63.89
BP	29.3306	4.283	24.5403	63.89
CS-BP	28.6662	3.9829	23.341	66.67
EEMD-BP	19.9351	2.6515	15.3561	86.11
EEMD-SE-BP	18.4103	2.6324	15.1279	75.00
EEMD-SE-CS-BP	14.2072	2.0932	11.9598	88.89
EEMD-VMD-SE-BP	12.3222	1.7905	10.3505	88.89
EEMD-VMD-SE-CS-BP	6.3521	0.8854	5.1545	97.22

表 4.12 广州白云国际机场模型预测结果比较(5 步预测)

模型	RMSE	MAPE(%)	MAE	D_{stat} (%)
ARIMA	38.0638	5.7680	32.6425	55.56
BP	33.0503	4.7938	28.1559	58.33
CS-BP	30.4111	4.2753	24.0850	66.67
EEMD-BP	21.0883	2.8190	16.4147	77.78
EEMD-SE-BP	19.3398	2.6426	15.4095	83.33
EEMD-SE-CS-BP	18.3383	2.5252	14.4237	86.11
EEMD-VMD-SE-BP	17.4307	2.3500	13.8146	83.33
EEMD-VMD-SE-CS-BP	7.9383	1.0411	5.8816	94.44

表 4.13 青岛胶东国际机场模型预测结果比较(1 步预测)

模型	RMSE	MAPE(%)	MAE	D_{stat} (%)
ARIMA	17.7997	5.9707	12.9235	61.11
BP	17.3879	6.9295	14.0776	55.56
CS-BP	15.1353	6.1271	12.6281	58.33
EEMD-BP	10.8977	4.2828	8.8087	72.22
EEMD-SE-BP	9.4649	3.4536	7.1860	83.33
EEMD-SE-CS-BP	8.2991	3.0276	6.3462	80.56
EEMD-VMD-SE-BP	6.8025	2.6499	5.1968	88.89
EEMD-VMD-SE-CS-BP	3.6079	1.4695	2.9658	86.11

表 4.14 青岛胶东国际机场模型预测结果比较(3 步预测)

模型	RMSE	MAPE(%)	MAE	D_{stat} (%)
ARIMA	18.5572	7.9535	15.8411	63.89
BP	15.5715	6.3609	12.6915	50.00
CS-BP	14.2930	5.5051	11.1602	75.00
EEMD-BP	13.6932	5.3557	10.8380	75.00
EEMD-SE-BP	9.6055	3.5535	7.4493	72.22
EEMD-SE-CS-BP	8.6692	3.6614	7.3912	83.33
EEMD-VMD-SE-BP	7.4418	2.9932	6.1169	83.33
EEMD-VMD-SE-CS-BP	6.3107	2.4242	4.9822	86.11

表 4.15 青岛胶东国际机场模型预测结果比较(5 步预测)

模型	RMSE	MAPE(%)	MAE	D_{stat} (%)
ARIMA	21.9134	7.7724	16.7793	50.00
BP	17.0672	7.1438	14.6758	69.44
CS-BP	17.3860	6.7616	13.4648	72.22
EEMD-BP	14.0888	4.9376	10.5381	69.44
EEMD-SE-BP	11.9223	4.9434	9.9958	66.67
EEMD-SE-CS-BP	10.7951	4.4731	9.0668	63.89
EEMD-VMD-SE-BP	9.2180	3.9039	7.9342	72.22
EEMD-VMD-SE-CS-BP	8.3390	3.1825	6.5489	83.33

通过分析不同模型的预测结果，我们可以得到以下结论：

(1)根据表 4.10-表 4.15 可知，与单一传统预测方法 ARIMA 模型相比，不论是广州白云国际机场还是青岛胶东国际机场的预测结果，神经网络 BP 模型、混合预测模型 CS-BP 以及基于分解-集成框架下的预测模型都具有更小的 MAPE、

RMSE、MAE，并且方向精度 D_{stat} 更高，说明对于非平稳数据，神经网络预测模型、分解-集成框架下的预测方法精度更高。

(2)比较添加智能优化算后的神经网络预测模型 BP、CS-BP，EEMD-SE-BP、EEMD-SE-CS-BP，EEMD-VMD-SE-BP、EEMD-VMD-SE-CS-BP，由表 4.10-表 4.15，广州白云国际机场一步预测模型，在这几个模型中，CS-BP 预测结果在 RMSE、MAPE、MAE 水平精度方面，较 BP 模型分别提高 40.85%、50.16%、49%，EEMD-SE-CS-BP 较 EEMD-SE-BP 分别提高 12.32%、12.33%、11.69%，且方向精度也稳步提升，智能优化算法使神经网络的参数更快适应数据，从而提高模型精度。

(3)EEMD-SE -BP 与 EEMD-BP 相比较，重构后的模型水平精度和方向精度均较高，青岛胶东国际机场的 RMSE 提升 45.54%，重构序列将频率相差不大的整合在一起，减少了分解集成预测时因分解个数过多而造成误差累积。

(4) EEMD-SE-CS-BP 与 EEMD-VMD-SE-CS-BP 相比较，文章提出的基于机场客流预测的二次分解集成模型 EEMD-VMD-SE-CS-BP 为最优模型，相较所有的基准模型，文章提出的模型仅在 MAPE 的精度分别提升了 85.36%、83.61%、82.47%、75.24%、74.21%、71.01%、67.53%，41.82%，此处仅列广州白云国际机场的相关指标，所以文章提出的基于二次分解模型 ICEEMDAN-VMD-SE-CS-BP 在预测机场客流方面由更好的表现。

(5)观察广州白云国际机场一步、三步、五步预测精度，比较本文提出的最佳二次分解模型 EEMD-VMD-SE-CS-BP，三步较一步，RMSE 精度降低 33.09%，五步预测比三步预测降低约 25%，时间间隔越久，预测的精度往往会更差。

5 分解集成框架下基于网络搜索信息的航空客流预测

基于机场自身的旅客吞吐量信息或者一些宏观因素进行建模，预测的信息来源或较为单一，或宏观指标信息不易及时获取。事实上，随着互联网时代的到来，越来越多的网民通过电脑、手机等客户端来获取信息。人们常常利用互联网搜索引擎来搜索自己所关注的事物或相关需求。根据一段时间内相关信息搜索量的变化，可以间接反映人们未来的行为特征。因此，本文将互联网搜索信息纳入到预测模型。在一些旅游型城市，游客流量与机场旅客吞吐量直接关联，因此，有必要进一步扩展搜索关键词。

5.1 模型框架

基于“分解-重构-集成”的思想提出了一种结合网络搜索信息的组合预测新模型。首先通过提取与机场旅客吞吐量直接或间接相关的网络搜索关键词信息，形成旅客吞吐量的辅助预测信息。然后将原始序列分解为若干子序列，并对各子序列重构后结合网络搜索信息分别进行预测，最后对各子序列的预测值进行综合集成。图 5.1 为本文的总体预测框架流程图。具体的预测过程主要包括如下三个部分：

(1)关键词筛选与综合搜索指数合成

步骤 1 关键词的初选。选取与机场旅客吞吐量直接或间接相关的网络搜索关键词信息。本文根据百度提供的与机场相关的需求图谱确定直接关键词，以机场所在城市文化旅游相关的信息作为间接关键词，初步选取 n 个网络搜索关键词；

步骤 2 关键词的筛选。首先，计算 n 个关键词搜索量与机场旅客吞吐量的 MIV 值，进而计算 n 个关键词的相对贡献率，本文从初选关键词中选取相对贡献率累计超 85%的 $\mathrm{m}(<=n)$ 个关键词；其次，计算 m 个关键词搜索量与机场旅客吞吐量的时差相关系数，并剔除不同滞后期时差相关系数均小于 0.5 的关键词，最终得到与机场旅客吞吐量相关性较强的 $\mathrm{d}(<=n)$ 个关键词。

步骤 3 综合搜索指数的合成。确定上述 d 个关键词的搜索量与旅客吞吐量相关性的最佳滞后期数，并计算最佳滞后期内关键词搜索总量与机场旅客吞吐量的皮尔逊相关系数 ρ_i，利用公式(4-7)，最终构造出综合搜索指数序列 $\{S_t\}$。

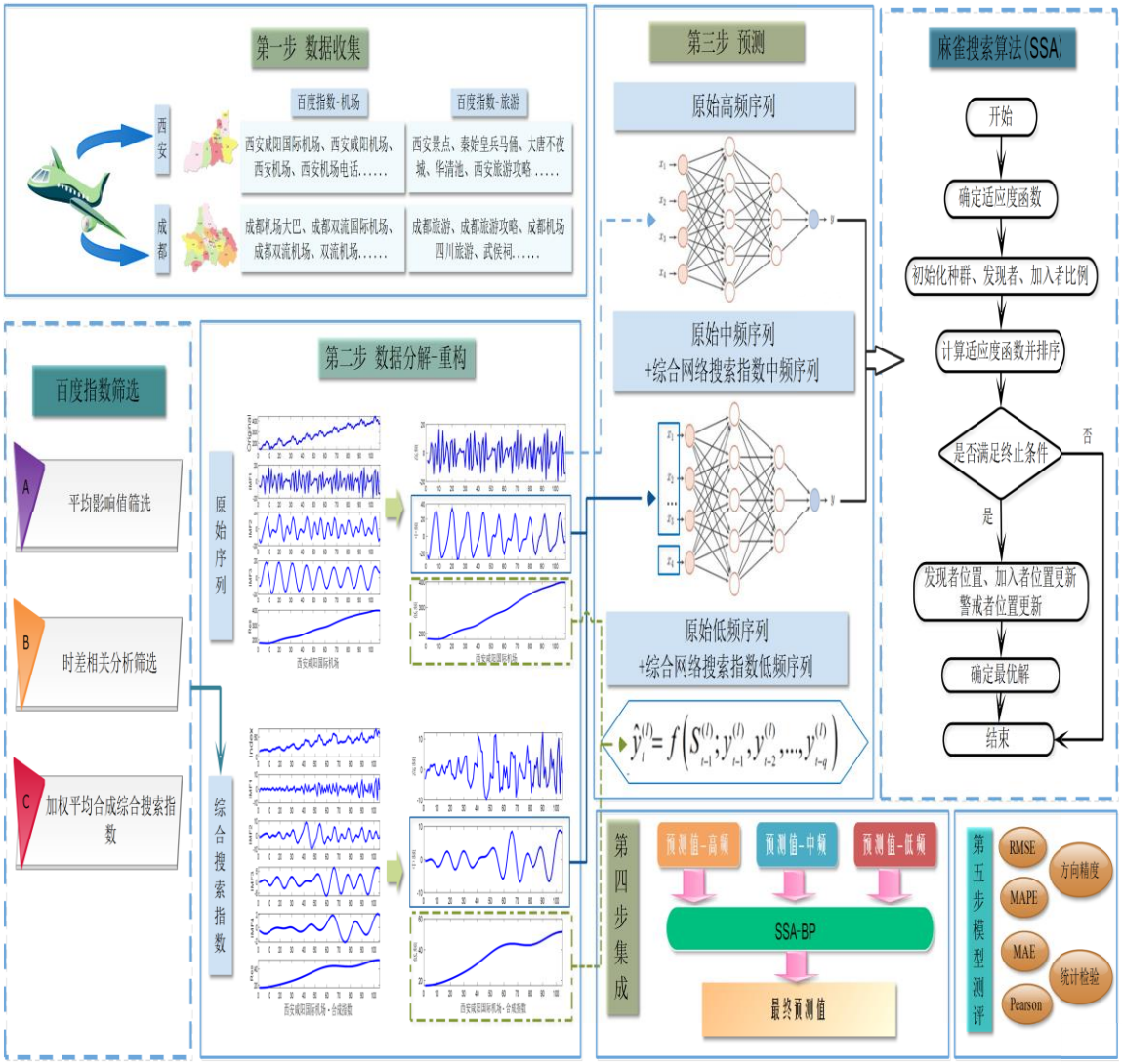

图5.1 总体框架流程图

(2)数据的分解与重构

步骤 4 数据的分解。利用 ICEEMDAN 分解方法分别将机场旅客吞吐量数据和综合搜索指数进行分解，得到旅客吞吐量数据的 n_1 个本征模态函数 IMF_1、…、IMF_{n1} 及残差 Res，以及搜索指数的 n_2 个本征模态函数 $\overline{IMF}_1$、…、$\overline{IMF}_{n2}$ 及残差 $\overline{Res}$。

步骤 5 子序列重构。分别计算旅客吞吐量数据及综合搜索指数分解之后各子模态序列和残差项的样本熵值，将样本熵值大于 1 的子序列合并成高频序列，样本熵值在 0 附近，小于 0.1 的合并为低频序列，其余合并为中频序列。

(3)数据的预测

步骤 6 子序列预测。本文参考梁小珍等[34]的处理方法，将综合搜索指数的高频项视为白噪声进行剔除。用搜索指数的中频(低频)信息辅助预测旅客吞吐量

的中频(低频)信息，利用试错法确定各序列自身的最佳输入维数。其中，对于高频和中频序列，由于非线性强、复杂度较高，采用基于 SSA 算法优化的 BP 神经网络来预测。低频项反映了原序列的长期变化趋势，因此本文采用传统的自回归分布滞后模型(ARDL)来预测，即结合机场旅客吞吐量低频序列与搜索指数的低频序列建立模型。其中，仍采用试错法选取机场旅客吞吐量低频序列的最佳滞后输入维数作为模型的自回归阶数，根据搜索指数的合成过程，其最佳滞后期取 1 即可。具体预测公式如下：

$$\hat{y}_t^{(l)} = f\left(S_{t-1}^{(l)}; y_{t-1}^{(l)}, y_{t-2}^{(l)}, ..., y_{t-q}^{(l)}\right) = \hat{\alpha} + \hat{\beta} S_{t-1}^{(l)} + \hat{\gamma}_1 y_{t-1}^{(l)} + \hat{\gamma}_2 y_{t-2}^{(l)} + ... + \hat{\gamma}_q y_{t-q}^{(l)} \tag{5-1}$$

这里 $y_t^{(l)}$ 和 $S_t^{(l)}$ 分别为旅客吞吐量和综合搜索指数经分解重构后的低频序列，q 为最佳滞后期。

步骤 7 非线性集成预测。利用训练样本集，将上述高、中、低频子序列的样本内预测值作为输入变量，将机场旅客吞吐量的真实值作为输出变量，重新建立一个新的 SSA-BP 神经网络模型，最后利用该模型可得到机场旅客吞吐量的非线性集成预测值。

5.2 数据分析与评测标准

本文分别选取西安咸阳和成都双流国际机场的月度旅客吞吐量数据进行实证分析(数据来源于 Wind 数据库)，数据的变化趋势和波动特征如图 4.2 所示，相关统计量如表 4.1 所示，数据的平稳性检验如表 4.2 所示。结果显示两个机场的旅客吞吐量均为非平稳的时间序列。两个机场的数据均涵盖了 2011 年 1 月至 2019 年 12 月共 108 个观测值，本文将 2011 年 1 月至 2017 年 12 月共 84 个观测值作训练集，2018 年 1 月至 2019 年 12 月共 24 个观测值作测试集。

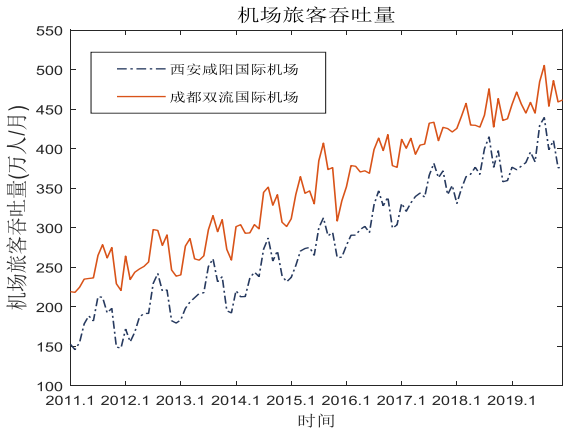

图5.2 机场旅客吞吐量

表 5.1 描述性统计量

序列 \ 统计量	均值	最大值	最小值	中位数	标准差	峰度	偏度
成都双流国际机场	350.46	505.47	218.45	349.19	79.64	1.80	0.01
西安咸阳国际机场	281.1	439.54	145.9	274.59	78.23	1.80	0.07

表 5.2 ADF 单位根检验——平稳性检验

	西安咸阳国际机场		成都双流国际机场	
	t 统计量	p 值	t 统计量	p 值
ADF 检验统计量	0.3793	0.9811	0.5094	0.9863
1%level	-3.5007		-3.4999	
5%level	-2.8922		-2.8919	
10%level	-2.5832		-2.5830	

为了更全面、客观地评估模型的预测性能，本文采用均方根误差(RMSE)、平均绝对百分比误差(MAPE)、平均绝对误差(MAE)三个水平指标，方向指标 D_{stat} 以及相关系数来度量模型的预测效果。具体公式见 3.2 中公式(3-1)-(3-4)。

$$r=\frac{\sum_{i=1}^{n}(y_i-\bar{y}_i)(\hat{y}_i-\bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}}_i)^2}} \tag{5-2}$$

其中 $\hat{y}_i$ 表示预测值，y_i 为实际值，$\bar{y}_i$ 表示实际值 y_i 的均值，$\bar{\hat{y}}_i$ 为预测值 $\hat{y}_i$ 的均值，n 表示观测样本的数目。进一步，为了从统计角度来比较不同模型的预测能力是否有显著性差异，本文引入 Diebold-Mariano(DM)检验方法。其中 DM 检验统计量为：

$$DM=\frac{E_d}{\sqrt{\hat{V}/n}} \tag{5-3}$$

这里 $d_i=(y_i-\hat{y}_{A,i})^2-(y_i-\hat{y}_{B,i})^2$ ，$E_d=\sum_{i=1}^{n}d_i/n$ ，$\hat{V}=\gamma_0+2\sum_{j=1}^{\infty}\gamma_j$ ，$\gamma_j=\mathrm{cov}(d_i,d_{i-j})$，$\hat{y}_{A,i}$ 表示 A 模型的预测值，$\hat{y}_{B,i}$ 表示 B 模型的预测值。α 为显著性水平，当 $|DM|>Z_\alpha$ 时拒绝原假设(Z_α 为正态分布单侧检验的临界值)，表明模型 A 优于模型 B。

5.2.1 关键词选取

(1) 关键词确定

成都双流国际机场与西安咸阳国际机场所在的成都和西安市都是我国的旅游热点城市，其大小景点数量近百处。人们旅游通常要查询吃住及旅游攻略，相关关键词网络搜索量的增加意味着潜在的游客流量增加，因而机场客流量也可能会随之增加。因此本文不仅选取与机场直接相关联的关键词，另外还选取与对应城市景点相关的关键词作为搜索信息。最后确定西安咸阳国际机场旅客吞吐量有关的关键词为：西安咸阳国际机场、咸阳机场电话、秦始皇兵马俑、大唐不夜城等 29 个关键词。与成都双流国际机场旅客吞吐量有关的关键词有：成都机场、成都机场大巴、成都双流机场、成都双流国际机场、成都景点、春熙路、锦里等 30 个关键词。上述关键词的搜索信息均为日度数据。

(2)关键词的筛选

为了与月度机场旅客吞吐量数据口径一致，我们将日度关键词搜索量整合为月度数据。以西安咸阳国际机场为例，首先分别计算 29 个关键词搜索量与机场旅客吞吐量数据的 MIV 值，得到每个关键词的相对贡献率，选取累积相对贡献率在 85%左右的关键词，初选得到与西安咸阳国际机场旅客吞吐量相关的搜索关键词共有 13 个。其次，由于旅客对旅游景点、机场相关关键词的搜索行为一般不会提前三个月，故计算出每个关键词滞后 1-3 期的搜索量与机场旅客吞吐量之间的时差相关系数并确定最佳的滞后期数，详细结果如表 5.3 所示。成都双流国际机场与搜索关键词的具体筛选结果如表 5.4 所示。然后基于相关系数的大小再一次筛选关键词，剔除与机场旅吞吐量的时差相关系数均小于 0.5 的关键词。由于西安旅游、西安华山旅游、咸阳国际机场对应的时差相关系数均小于 0.5，故剔除上述 3 个关键词。同理，成都双流国际机场剔除了成都双流机场、成都机场大巴时刻表、成都旅游景点大全 3 个关键词。最终，确定与西安咸阳国际机场相关度较高的关键词为 10 个，成都双流国际机场为 14 个。

表 5.3 西安咸阳国际机场关键词累积相对贡献率与时差相关系数

序号	关键词	相对贡献率(%)	累积相对贡献率(%)	时差相关系数			
				滞后 1 期	滞后 2 期	滞后 3 期	最佳滞后期
1	西安机场大巴时刻表	9.83	9.83	0.63	0.65	0.65	3
2	大唐不夜城	8.78	18.61	0.58	0.64	0.63	2
3	西安旅游	8.21	26.82	0.39	0.35	0.34	1

续表 5.3 西安咸阳国际机场关键词累积相对贡献率与时差相关系数

4	西安华山旅游	7.06	33.88	0.14	0.09	0.04	1
5	西安自助游	6.95	40.84	0.54	0.61	0.65	3
6	陕西历史博物馆	6.32	47.15	0.84	0.83	0.81	1
7	西安美食攻略	6.20	53.35	0.68	0.65	0.63	1
8	西安旅游攻略	6.20	59.55	0.57	0.53	0.52	1
9	咸阳国际机场	5.88	65.43	0.16	0.14	0.10	1
10	华清池	5.74	71.17	0.59	0.57	0.55	1
11	大雁塔	4.81	75.97	0.61	0.60	0.78	3
12	西安景点	4.20	80.17	0.85	0.81	0.77	1
13	西安回民街	3.29	83.46	0.79	0.76	0.75	1

表 5.4 成都双流国际机场关键词累积相对贡献率与时差相关系数

序号	关键词	相对贡献率(%)	累积相对贡献率(%)	时差相关系数			
				滞后 1 期	滞后 2 期	滞后 3 期	最佳滞后期
1	春熙路	10.55	10.55	0.53	0.52	0.50	1
2	成都旅游攻略	9.77	20.32	0.54	0.53	0.50	1
3	望江楼	7.02	27.34	0.60	0.62	0.63	3
4	成都旅游景点	6.98	34.32	0.51	0.48	0.46	1
5	双流机场大巴	5.56	39.88	0.74	0.75	0.74	2
6	成都大熊猫基地	5.10	44.98	0.82	0.81	0.81	1
7	宽窄巷子	5.06	50.04	0.84	0.84	0.83	2
8	青羊宫	4.54	54.58	0.57	0.56	0.55	1
9	成都双流机场	3.89	58.47	0.48	0.46	0.46	1
10	锦里	3.79	62.25	0.75	0.74	0.71	1
11	双流机场	3.72	65.97	0.52	0.51	0.51	1
12	成都景点	3.63	69.60	0.88	0.84	0.86	1
13	成都机场大巴时刻表	3.44	73.04	0.01	0.01	0.03	3
14	成都旅游景点大全	2.88	75.92	0.10	0.11	0.10	2
15	成都天气	2.78	78.70	0.72	0.71	0.71	1
16	成都旅游	2.68	81.38	0.45	0.43	0.42	1
17	文殊院	2.55	83.93	0.88	0.88	0.88	2

(3) 综合搜索指数合成

根据筛选得到的关键词最佳期滞后期数对每个关键词在最佳滞后期内的搜索量进行累加，计算其与机场旅客吞吐量的皮尔逊相关系数作为权重，通过加权平均合成综合搜索指数。成都双流国际机场与西安咸阳机场的综合搜索指数与其机场旅客吞吐量的关系如图 5.3 所示。通过表 5.4 的比较发现，本文提出的关键词合成方法与关键词搜索量简单相加、基于皮尔逊相关系数的加权相加以及关键词最佳滞后期内搜索量的简单相加相比，本文合成的综合搜索指数与机场旅客吞吐量的相关系数最高。通过对搜索指数与对应机场旅客吞吐量进行格兰杰因果检验，如表 4.6 所示，得到 P 值均小于 0.05，说明搜索指数与机场旅客吞吐量存在格兰杰因果关系，即存在长期均衡关系。

表 5.5 综合搜索指数与机场旅客吞吐量相关系数

	简单加总	加权加总	最佳滞后期累计加总	权重、滞后累计加总
西安咸阳国际机场	0.8005	0.8293	0.8472	0.8699
成都双流国际机场	0.7282	0.7308	0.6944	0.7586

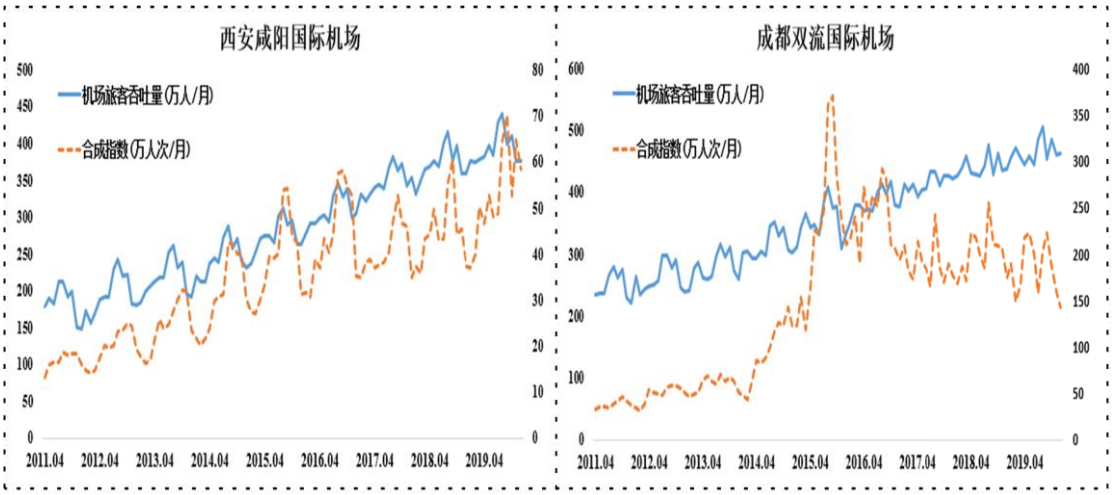

图5.3 综合搜索指数与机场旅客吞吐量比较图

表 5.6 搜索指数与机场旅客吞吐量的格兰杰因果检验

原假设	F 统计量	P 值	
搜索指数不是机场旅客吞吐量的格兰杰原因(西安)	2.7288	0.0338	拒绝原假设
搜索指数不是机场旅客吞吐量的格兰杰原因(成都)	3.3155	0.0404	拒绝原假设

5.2.2 数据的分解-重构-集成过程

(1)数据分解与重构

利用 ICEEMDAN 方法将机场旅客吞吐量数据与综合搜索指数分别进行分解，其中成都双流国际机场、西安咸阳国际机场的旅客吞吐量数据均被分解为 3 个本征模态函数 IMFs 与 1 个残差趋势项 Res。双流国际机场和咸阳国际机场的百度综合搜索指数分别被分解为 6 个和 5 个子序列。由于样本熵值的大小可反映序列的复杂度，通过计算上述各子序列的样本熵值，本文将样本熵值大于 1 的子模态序列合并为高频序列，样本熵值在 0 附近的子序列合并为低频序列，其余子序列合并为中频序列。由于成都双流国际机场的搜索指数经分解后的子序列其样本熵值均小于 1，故仅重构成中、低频序列。图 5.4 显示了两个机场旅客吞吐量分解后的子序列及样本熵值，图 5.5 显示了重构后的序列及其样本熵值。

(2)子序列预测

类似文献[33]的处理方法，将搜索指数分解得到的高频项视为噪声进行剔除，不参与预测。首先，机场旅客吞吐量的高频序列采用经麻雀搜索算法优化的 BP 神经网络模型(SSA-BP)来预测。为了进一步提高预测精度，本文采用试错法来确定每个序列的最佳滞后期数，即神经网络输入层的维数。本文尝试采用滞后 2~10 期的数据作为输入来建立预测模型，使用 RMSE 作为评测标准来判定不同滞后期数的优劣。其次，将搜索指数的中频序列作为辅助输入信息，通过建立 SSA-BP 模型对机场旅客吞吐量中频序列进行预测，同样采用试错法确定机场旅客吞吐量的最佳滞后期数，而搜索指数根据定义直接使用滞后 1 期作为输入变量。由此，机场旅客吞吐量中频序列的最佳滞后期数加上搜索指数的滞后 1 期即为 SSA-BP 模型输入维数。图 4.6 显示西安咸阳国际机场进行一步预测时高频和中频序列对应的最佳滞后期数分别为 7 和 9，而成都双流国际机场的高频和中频序列最佳滞后期均为 10。最后，注意到机场旅客吞吐量的低频序列复杂度低，趋势性强，建立基于自身滞后序列和搜索指数低频序列滞后 1 期为影响因素的 ARDL 模型，通过比较预测结果的 RMSE 来确定自回归部分的最优滞后阶数，最后确定西安咸阳国际机场和成都双流国际机场低频序列的最优滞后期分别为 4 和 10。

(3)综合集成

将上述(2)中得到的高、中、低频子序列的预测值作为输入变量，机场旅客吞吐量对应预测期的真实值作为输出变量，重新建立 SSA-BP 模型，通过对高、中、低频序列的非线性集成，得到最终的预测值。

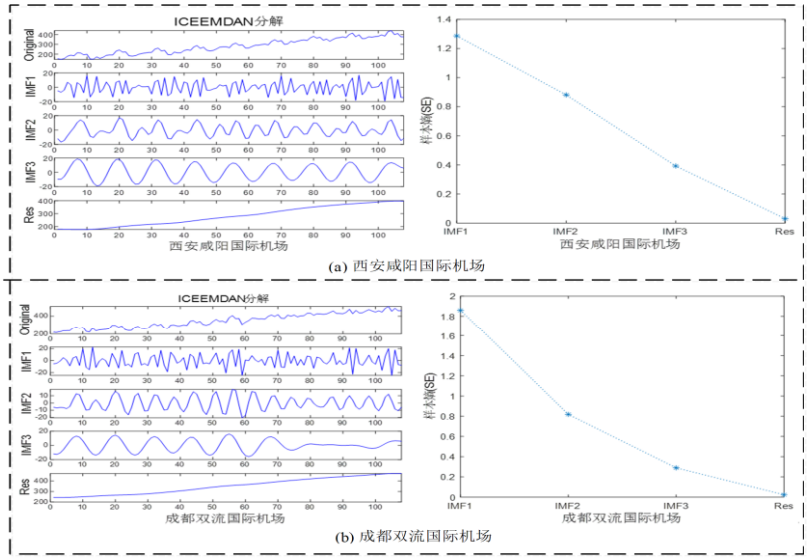

图5.4 机场旅客吞吐量的分解及其子序列样本熵值

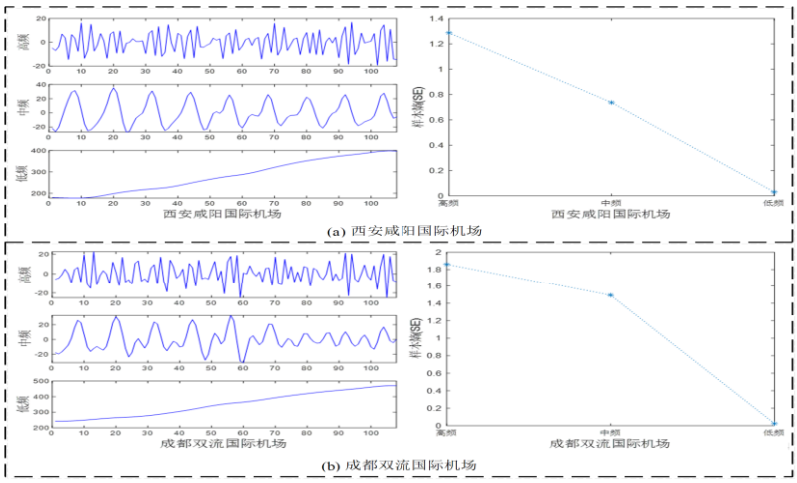

图5.5 机场旅客吞吐量重构序列及其样本熵值

5.2.3 结果分析

为了评估在本文预测框架下所建模型的预测能力，本节对多个模型的预测效果进行了比较。首先，为了考查网络搜索信息的辅助预测能力，本文将未加入和加入网络搜索信息的单一预测模型以及组合预测模型分别进行考察；其次，为了体现组合预测模型的预测优势，与单一预测模型进行了比较；最后，为了评估模型的稳健性，对机场旅客吞吐量的一、二、三步预测效果分别进行了比较。结果如表 5.7-表 5.14 和图 5.7-图 5.8 所示。其中，本文提出的基于网络搜索信息的”分解-重构-集成”组合预测新模型标记为 ICE-SE-SSA-BP-AR (index)，这里”ICE”表示用 ICEEMDAN 进行分解，”SE”代表重构，”SSA-BP-AR”表示预测方法采用了 SSA 算法优化的 BP 神经网络和自回归分布滞后模型，”index”表示模型中加入了网络搜索信息。SSA-BP(index)和 BP(index)分别表示加入网络搜索信息后经

SSA 算法优化和未经优化的 BP 神经网络模型；而未加入网络搜索信息的分解-重构-集成模型、分解-集成模型、经 SSA 算法优化的 BP 神经网络模型以及未经优化的 BP 神经网络模型分别标记为：ICE-SE-SSA-BP、ICE-SSA-BP、SSA-BP 和 BP。

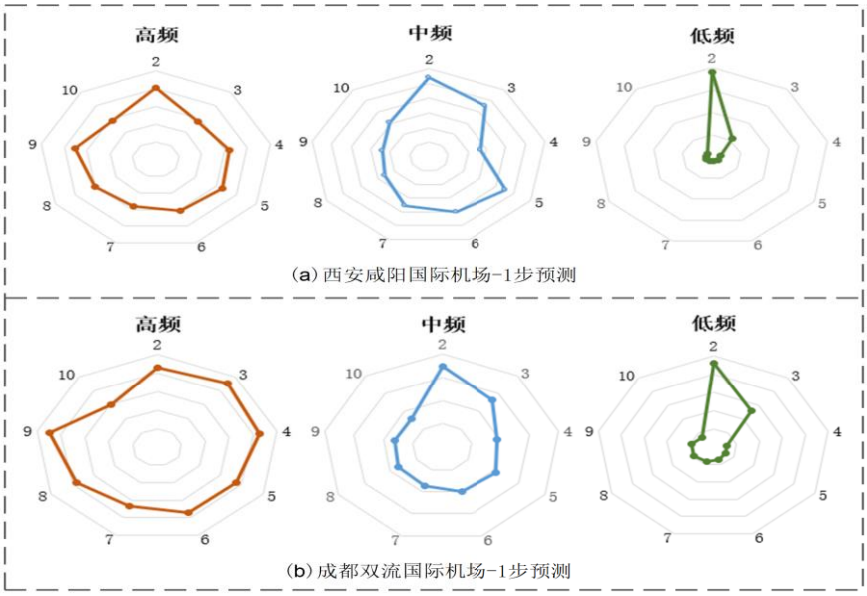

图5.6 高、中、低频序列不同滞后期对应的均方根误差(RMSE)

表 5.7 西安咸阳国际机场模型预测结果比较(1 步预测)

模型	RMSE	MAPE(%)	MAE	D_{stat} (%)	Person
ICE-SE-SSA-BP-AR (index)	5.87	1.23	4.66	91.67	0.9759
ICE-SE-SSA-BP	6.27	1.32	5.12	87.50	0.9792
ICE-SSA-BP	8.26	1.77	6.84	83.33	0.9682
SSA-BP(index)	15.36	3.09	11.62	87.50	0.8254
SSA-BP	15.74	3.09	11.90	75.00	0.7641
BP(index)	17.55	3.61	13.67	79.17	0.7373
BP	18.00	3.49	13.47	70.83	0.7373

表 5.8 西安咸阳国际机场模型预测结果比较(2 步预测)

模型	RMSE	MAPE(%)	MAE	D_{stat} (%)	Person
ICE-SE-SSA-BP-AR (index)	6.59	1.28	4.90	91.67	0.9651
ICE-SE-SSA-BP	7.28	1.56	5.90	91.67	0.9545
ICE-SSA-BP	8.71	1.84	7.06	83.33	0.9381
SSA-BP(index)	15.71	3.15	11.91	70.83	0.7793
SSA-BP	16.99	3.22	12.42	70.83	0.7439

续表 5.8 西安咸阳国际机场模型预测结果比较(2 步预测)

BP(index)	17.90	3.59	13.45	79.17	0.7375
BP	19.02	3.94	15.02	70.83	0.6449

表 5.9 西安咸阳国际机场模型预测结果比较(3 步预测)

模型	RMSE	MAPE(%)	MAE	D_{stat} (%)	Person
ICE-SE-SSA-BP-AR (index)	7.11	1.49	5.62	91.67	0.9612
ICE-SE-SSA-BP	8.71	1.76	6.63	87.50	0.9412
ICE-SSA-BP	9.09	1.89	7.24	87.50	0.9474
SSA-BP(index)	17.62	3.68	13.92	83.33	0.8426
SSA-BP	18.13	3.85	14.72	75.00	0.7105
BP(index)	18.41	4.14	15.46	75.00	0.7803
BP	19.56	4.21	16.00	75.00	0.7147

表 5.10 成都双流国际机场模型预测结果比较(1 步预测)

模型	RMSE	MAPE(%)	MAE	D_{stat} (%)	Person
ICE-SE-SSA-BP-AR (index)	5.76	1.02	4.61	91.67	0.9614
ICE-SE-SSA-BP	6.98	1.08	5.00	87.50	0.9555
ICE-SSA-BP	9.05	1.47	6.83	83.33	0.9423
SSA-BP(index)	17.12	3.01	13.73	66.67	0.5775
SSA-BP	17.28	3.21	14.39	62.50	0.5912
BP(index)	17.54	3.36	15.33	62.50	0.5769
BP	18.64	3.40	15.43	70.83	0.5087

表 5.11 成都双流国际机场模型预测结果比较(2 步预测)

模型	RMSE	MAPE(%)	MAE	D_{stat} (%)	Person
ICE-SE-SSA-BP-AR(index)	12.37	2.12	9.80	79.17	0.8449
ICE-SE-SSA-BP	13.05	2.33	10.64	75.00	0.7722
ICE-SSA-BP	16.90	2.93	13.39	75.00	0.5754
SSA-BP(index)	18.45	2.90	13.48	75.00	0.6075
SSA-BP	18.05	3.06	13.97	79.17	0.5200
BP(index)	18.84	3.39	15.45	75.00	0.4262
BP	20.84	3.94	17.84	66.67	0.3342

表 5.12 成都双流国际机场模型预测结果比较(3 步预测)

模型	RMSE	MAPE(%)	MAE	D_{stat} (%)	Person

续表 5.12 成都双流国际机场模型预测结果比较(3 步预测)

ICE-SE-SSA-BP-AR (index)	14.44	2.73	12.36	79.17	0.7735
ICE-SE-SSA-BP	15.02	2.79	12.66	79.17	0.6979
ICE-SSA-BP	16.20	2.73	12.66	79.17	0.7668
SSA-BP(index)	17.04	2.94	13.28	75.00	0.5969
SSA-BP	18.73	3.69	16.66	66.67	0.6063
BP(index)	18.97	3.38	15.42	70.83	0.4307
BP	21.08	3.68	17.02	66.67	0.3372

表 5.13 西安咸阳国际机场 DM 检验

模型	1 步		2 步		3 步	
	DM	P 值	DM	P 值	DM	P 值
ICE-SE-SSA-BP-AR (index)	-	-	-	-	-	-
ICE-SE-SSA-BP	-0.60	0.2741	-0.87	0.1934	-0.75	0.2265
ICE-SSA-BP	-2.62***	0.0044	-2.88***	0.0020	-2.13**	0.0168
SSA-BP(index)	-2.68***	0.0036	-3.10***	0.0010	-3.46***	0.0003
SSA-BP	-3.10***	0.0010	-2.54***	0.0055	-3.34***	0.0004
BP(index)	-2.67***	0.0038	-2.81***	0.0025	-4.18***	0.0000
BP	-2.86***	0.0021	-3.19***	0.0007	-3.56***	0.0002

表 5.14 成都双流国际机场 DM 检验

模型	1 步		2 步		3 步	
	DM	P 值	DM	P 值	DM	P 值
ICE-SE-SSA-BP-AR (index)	-	-	-	-	-	-
ICE-SE-SSA-BP	-1.04	0.1494	-0.38	0.3512	-0.35	0.3638
ICE-SSA-BP	-1.75**	0.0397	-1.56**	0.0597	-0.59	0.2780
SSA-BP(index)	-3.34***	0.0004	-1.45**	0.0742	-0.95	0.1698
SSA-BP	-3.82***	0.0001	-1.61*	0.0539	-1.69**	0.0460
BP(index)	-4.12***	0.0000	-2.76***	0.0028	-1.23	0.1092
BP	-4.04***	0.0000	-3.06***	0.0011	-1.58*	0.0574

***表示1%置信水平；**表示5%置信水平;*表示10%置信水平

通过分析表 5.7-5.14 不同模型的预测结果，我们可以得到以下结论：

(1)对于 BP、SSA-BP、ICEEMDAN-SSA-BP 及 ICEEMDAN-SE-SSA-BP 四种未考虑网络搜索信息的预测模型而言，基于分解-重构-集成框架下的预测模型，在 RMSE、MAPE、MAE 三个指标下均得到最小值，说明 ICEEMDAN-SE-SSA-BP 通过分解降低了序列的复杂度，同时重构避免了分解过

多子序列造成误差累积，提高了模型的精度。

(2)不论是单一预测模型还是分解-重构-集成预测模型，加入网络搜索信息的模型在各指标上都取得了比未加入搜索信息的模型更高的预测精度。以西安咸阳机场的 1 步预测为例，采用 ICE-SE-SSA-BP-AR(index)模型后要比 ICE-SE-SSA-BP 模型的 RMSE、MAPE、MAE 值分别减少了 6.38%、6.82%、8.98%，而成都双流机场 1 步预测所对应的评测指标分别减少了 17.48%、5.56%、7.80%，同时方向预测精度也明显提高，说明在本文预测框架下加入网络搜索辅助信息可以提高在单一原始序列信息建模下的预测精度。

(3) 为了检验基于网络搜索信息的分解-重构-集成预测模型的稳健性，采用多步预测进行进一步比较，发现网络搜索信息始终具有良好的辅助预测能力。和其他基准模型相比，本文提出的基于网络搜索信息的分解-重构-集成预测模型在两个机场旅客吞吐量的 1、2、3 步预测中，都显示出明显的优势，并且由图 5.7-5.8 可以看出，预测步长越短，模型的预测精度越高，1 步预测比 2、3 步预测有更高的预测精度，采用基于搜索信息的 ICEEMDAN-SE-SSA-BP 模型在西安咸阳国际机场的 1 步预测比 2、3 步预测的 MAPE 值分别减少了 3.91%和 17.45%，成都双流国际机场 1 步预测比 2、3 步预测的 MAPE 值分别减少了 51.89%和 62.64%。

通过表 5.13-5.14 的 DM 检验得到，除 ICEEMDAN-SE-SSA-BP 模型、成都双流国际机场 ICEEMDAN-SSA-BP、SSA-BP(index)与 BP(index)模型的三步预测外，本文提出的基于网络搜索信息的分解-重构-集成模型明显优于其余模型。

图 5.7 西安咸阳国际机场 1、2、3 步预测误差比较

图5.8 成都双流国际机场1、2、3步预测误差比较

6 结语

本文主要利用”分解-集成”的思想，对我国港口集装箱吞吐量、机场旅客吞吐量进行非线性预测。一方面，从预测效果上来看，不论是采用一次、二次分解-集成框架下的组合预测模型、基于网络搜索信息的分解-集成预测模型，都要比单一传统时序预测、神经网络预测模型预测模型具有更好的预测效果。主要原因在于：其一，对于非线性、高波动的复杂数据而言，利用分解算法能得到了不同频率的子序列，从而提高预测精度。其二，不论是对一次分解、二次分解得到的子序列进行重构，目的在于削减了预测时的累积误差。其三，结合粒子群优化算法、布谷鸟优化算法、麻雀搜索算法优化神经网络更进一步提高了模型的精度。另一方面，本文在”分解-重构-集成”思想的基础上，结合机场旅客吞吐量相关的网络搜索关键词信息，提出了一种基于网络搜索信息的分解-集成预测模型，是通过增加不同维度的数据辅助预测。通过分解降低序列的复杂度，重构减少误差累积，将网络搜索信息引入模型，从不同维度着手，增加了模型预测精度。

本文提出的预测方法为时间序列的的准确预测提供了参考，但本文仍存在一些不足，譬如，在原始数据的分解过程中，本文与多数文献做法一样，仅为了便于比较不同模型的优劣，直接将训练和测试样本进行了一次性分解，但在实际应用中，测试样本是未知的，此时需要参考 Xiao[48]等提出的方法，仅基于训练样本进行分解再逐步预测和集成，这将是我们未来研究的工作。

参考文献

[1]刘斌,邱国栋,刘超.集装箱港口吞吐量线性回归模型[J].大连海事大学学报,2003(02):27-30

[2]张丽,闫世锋.Holt-Winters 方法与 ARIMA 模型在中国航空旅客运输量预测中的比较研究[J].上海工程技术大学学报,2006(03):280-283.

[3]姚晏斌,高金华.灰色模型GM(1,2)在机场旅客吞吐量预测中的应用[J].中国民航飞行学院学报,2006(04):12-16.

[4]Ching-Fu Chen,Yu-Hern Chang,Yu-Wei Chang.Seasonal ARIMA forecasting of inbound air travel arrivals to Taiwan[J]. Transportmetrica,2009,5(2):125-140.

[5]李明捷.基于三次指数平滑法的航空旅客运输量预测[J].中国民航飞行学院学报,2009,20(01):43-44+47.

[6]Erma,Suryani.Air passenger demand forecasting and passenger terminal capacity expansion:A system dynamics framework[J].Expert Systems with Applications,2010,37(3):2324-2339.

[7]黄邦菊,林俊松,郑潇雨,方学东.基于多元线性回归分析的民用运输机场旅客吞吐量预测[J].数学的实践与认识,2013,43(04):172-178.

[8]Tsui W,Balli HO ,Gilbey A,Gow H.Forecasting of Hong Kong airport's passenger throughput[J].Tourism Management,2014,42(42):62–76.

[9]Seongdo Kim,Do Hyoung Shin. Forecasting short-term air passenger demand using big data from search engine queries[J].Automation in Construction,2016(70):98-108.

[10]王婷婷,王紫晨.基于灰色马尔科夫对龙洞堡机场旅客吞吐量预测[J].交通科技与经济,2017,19(05):48-51.

[11]杨梦达,戴晨斌,冀和,樊重俊.基于 SD-ARIMA 复合模型的虹桥机场旅客吞吐量预测[J].物流科技,2019,42(11):74-77+87.

[12]刘长俭,张庆年.基于时间序列 BP 神经网络的集装箱吞吐量动态预测[J].水运工程,2007(1):4-7+11.

[13]李季涛,马彩雯,孙光祈.基于 RBF 神经网络的港口集装箱吞吐量动态预测[J].大连交通大学学报,2008(04):27-32.

[14]吴慧军,刘桂云.基于灰色 RBF 组合模型的宁波港集装箱海铁联运量预测[J].

宁波大学学报(理工版),2016,29(04):123-127.

[15]李广儒,朱庆辉.基于 Elman 神经网络的港口货物吞吐量预测[J].重庆交通大学学报(自然科学版),2020,39(06):8-12.

[16]程文忠,任凤香,周宣赤.SVM 在九江港吞吐量预测中的应用研究[J].物流工程与管理,2012,34(09):61-64.

[17]肖海波,秦鲁敏,阳劲.基于 BP 神经网络的机场旅客吞吐量预测[J].山西科技,2005(05):121-122.

[18]冯兴杰,魏新,黄亚楼.基于支持向量回归的旅客吞吐量预测研究[J].计算机工程,2005(14):172-173.

[19]Yan K W.Study on the Forecast of Air Passenger Flow Based on SVM Regression Algorithm[C]//First International Workshop on Database Technology & Applications.IEEE Computer Society,2009.

[20]廖洪一,王欣.极限学习机在机场旅客吞吐量预测中的应用[J].计算机系统应用,2015,24(11):257-261.

[21]王子位.基于长短期记忆网络的民航流量预测方法[J].数字通信世界,2018(04):264.

[22]李洁,林永峰.基于多时间尺度 RNN 的时序数据预测[J].计算机应用与软件,2018,35(07):33-37+62.

[23]Shaolong Sun,Hongxu Lu,Kwok-Leung Tsui,Shouyang Wang. Nonlinear vector auto-regression neural network for forecasting air passenger flow[J]. Journal of Air Transport Management,2019(78):54-62.

[24]Korkmaz E,Akgüngör A P.The forecasting of air transport passenger demands in Turkey by using novel meta‐heuristic algorithms[J].Concurrency and Computation:Practice and Experience,2021,33(16):e6263.

[25]Li Long Chan,Guleria Yash,Alam Sameer.Air passenger forecasting using Neural Granger causal Google trend queries[J].Journal of Air Transport Management,2021(95):102083.

[26]赵尚威,周建红.中国港口集装箱吞吐量预测:基于组合时间序列[J].系统科学与数学,2018,38(02):210-219.

[27]蒋惠园,张安顺.组合预测模型在武汉港集装箱吞吐量预测中的应用[J].物流技术,2020,39(02):44-47+140.

[28]冯宏祥,GRIFOLL Manel,AGUSTI Martinmallofre.基于数据分解的上海港集装箱吞吐量预测模型[J].中国航海,2019,042(002):132-138.

[29]屈拓.组合模型在机场旅客吞吐量预测中的应用[J].计算机仿真,2012,29(04):108-111.

[30]Gang Xie,Shouyang Wang,Kin Keung Lai.Short-term forecasting of air passenger by using hybrid seasonal decomposition and least squares support vector regression approaches[J]. Journal of Air Transport Management,2014(3):20-26.

[31]梁小珍,乔晗,汪寿阳,张珣.基于奇异谱分析的我国航空客运量集成预测模型[J].系统工程理论与实践,2017,37(06):1479-1488.

[32]李苑辉,刘夏,欧志鹏.基于 ARIMA 模型的三亚机场客流量预测[J].软件,2018,39(07):42-47.

[33]Feng Jin,Yongwu Li,Shaolong Sun,Hongtao Li.Forecasting air passenger de-mand with a new hybrid ensemble approach[J].Journal of Air Transport Management,2020(83):101744.

[34]梁小珍,张晴,杨明歌.面向网络搜索数据的航空客运需求两阶段分解集成预测模型[J].管理评论,2021,33(05):236-245.

[35]Huang N E,Shen Z,Long S R.The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J].Proceedings of the Royal Society of London.Series A:mathematical,physical and engineering sciences,1998,454(1971):903-995.

[36]Wu Z H, Huang N E.Ensemble empirical mode decomposition:A noise-assisted data analysis method[J].Advances in Adaptive Data Analysis,2009,1(01):1-41.

[37]Yeh J R, Shieh J S, Huang N E.Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method[J].Advances in Adaptive Data Analysis,2010,2(02):135-156.

[38]Torres M E,Colomians M A,Schlotthuer G.Complete ensemble empirical mo de decomposition with adaptive noise[J].Brain Research Bulletin,2011,125(3): 4144-4147.

[39]Lin H,Sun Q,Chen S Q.Reducing Exchange Rate Risks in International Tra de:A Hybrid Forecasting Approach of CEEMDAN and Multilayer LSTM[J]. Sustainability,2020,12(6):2451.

[40]Niu H L, Xu K L, Liu C.A decomposition-ensemble model with regrouping

method and attention-based gated recurrent unit network for energy price prediction[J].Energy,2021(231):120941.

[41]Dragomiretskiy K,Zosso D. Variational mode decomposition[J].IEEE transactions on signal processing,2013,62(3): 531-544.

[42]Richman J S,Randall M J.Physiological time-series analysis using approximate entropy and sample entropy[J].American Journal of Physiology Heart & Circulatory Physiology,2000,278(6):H2039.

[43]Vapnik V.The nature of statistical learning theory[M].New York:Springer Verlag，1999.

[44]Kennedy J,Eberhart R.Particle swarm optimization[C]//Proceedings of ICNN'95-international conference on neural networks.IEEE,1995,4:1942-1948.

[45]Yang X S,Deb S.Cuckoo Search via Lévy flights[C]//World Congress on Nature&Biologically Inspired Computing.IEEE, 2009:210-214.

[46]Xue J, Shen B.A novel swarm intelligence optimization approach: sparrow search algorithm[J].Systems Science & Control Engineering,2020,8(1):22-34.

[47]Dombi G W,Nandi P,Saxe J M.Prediction of Rib Fracture Injury Outcome by an Artificial Neural Network[J].The Journal of trauma,1995,39(5):915-921.

[48]Xiao Y J,Wang X K,Wang J Q.An adaptive decomposition and ensemble model for short-term air pollutant concentration forecast using ICEEMDAN-ICA[J]. Technological Forecasting and Social Change,2021(166):120655.

致谢

写到这里，才发觉我人生中一个非常重要的阶段即将结束，我的研究生阶段即将画上句号。三年时间转瞬即逝，考研时的辛苦恍如昨日，拿到研究生录取通知书时的喜悦还清晰的印在脑海里。虽说我运气比较好，高考掉尾巴考上大学，压分考上研究生，在但是三年的研究生，确确实实有努力学习专业知识，也在了学业上取得进步，认识了新的朋友，可以说是收获颇多。这一切的成果离不开家人，老师和同学的支持。

首先，我要感谢我的老师，您严谨治学的态度和一丝不苟的负责精神永远值得我学习。帮我和同门找到一个前沿的方向，和我们一起学习，一起进步，要是没有您认真负责的督促我们不断学习，可能三年时间荒废度过。我犹记得，凌晨两点多您还将我没懂的知识详细解释，留言到微信。即使假期，写论文遇到瓶颈，您还能耐心打电话告诉我接下来如何进行，怎样进一步分析。论文的每一稿，您都自己打印出来逐字进行批改，看着上面密密麻麻的批注，真正的传道授业解惑便是如此吧。在读研究生的期间，能得到您的指导是最大的幸运。

其次，我要感谢父母，永远是我最温暖的避风港。在读书方面，你们给予了我最大的支持，这将将二十载的辛苦读书，你们都给我自由选择的权利，你们对我选择的尊重就是给我最大的鼓励。在日常生活中，妈妈永远是最温柔坚实的力量，关心我生活的点点滴滴，爸爸永远是在我遇到困难想放弃时，鼓励我再坚持一下，就会成功。在人生漫长的道路上，有你们的守护，我将会走得更远，更快乐。

然后，我要感谢师三位师妹，经常和你们讨论问题，让我受益匪浅，既是你们的师姐，更是你们的朋友，你们勤奋学习的劲头，更是催促我不懈努力的动力。感谢朝夕相处的室友，研究生生涯有你们真的很开心。你们平时细心贴心，学校就像家一样温暖。

最后，向所有关心、爱护和支持我的人表示由衷的感谢。感谢百忙之中评阅论文的老师，以及毕业答辩的老师。

硕士期间论文成果

1. 于婷,孙景云.基于 EEMD 和 PSO 方法的我国港口集装箱吞吐量预测[J].物流技术,2021,40(05):56-64.
2. 孙景云,于婷,何林芸. 基于网络搜索信息的多模态数据驱动航空客流集成预测[J].运筹与管理,已录用(2022.04.24).