

doi: 10.12012/CJoE2022-0110

多源异构数据驱动的后疫情时期旅游需求 预测方法研究

武静¹, 赵二龙¹, 孙少龙¹, 汪寿阳^{2,3,4}

(1. 西安交通大学管理学院, 西安 710049; 2. 中国科学院数学与系统科学研究院, 北京 100190; 3. 中国科学院预测科学研究中心, 北京 100190; 4. 上海科技大学创业与管理学院, 上海 201210)

摘要 准确的旅游需求预测对旅游业高质量发展有着重要的作用, 特别是中短期客流量预测对旅游目的地的旅游资源调配及应急管理至关重要, 新冠疫情的冲击导致旅游需求出现结构性变化, 将对旅游需求预测预警提出新的挑战。本研究主要聚焦后疫情时期多源异构数据融合是否能够提升旅游需求预测方法的性能, 为了探究该问题, 本研究首先从旅游者生成的在线评论、旅游者关注的网络搜索数据和旅游目的地实时在线新闻数据中提取众多影响旅游需求波动的变量, 并采用主题模型、情感分析及特征工程方法对其进行处理; 其次, 根据变量时间频度的不同采用混频建模的方法对变量进行融合处理得到预测的多模态融合特征; 最后, 基于多源异构数据融合驱动 SARIMA-MIDAS 预测方法对旅游需求进行建模预测。该研究工作主要以后疫情时期中国香港游客量建模预测为研究对象, 实证结果揭示出本研究提出的多源异构数据融合驱动 SARIMA-MIDAS 预测方法在后疫情期间能够取得最佳的预测表现, 因此, 该研究的结果可为旅游需求预测提供一种新的解决方案, 为相关政府机构和从业者提供决策支持。

关键词 旅游需求预测; 多源异构数据; 混频数据融合; 用户生成内容; 情感分析

收稿日期: 2022-11-29

基金项目: 国家重点研发计划项目 (2022YFF0903000); 国家自然科学基金 (72101197, 71988101); 陕西省软科学项目 (2023-CX-RKX-081)

Supported by National Key R&D Program of China (2022YFF0903000); National Natural Science Foundation of China (72101197, 71988101); Soft Science Research Program of Shaanxi (2023-CX-RKX-081)

作者简介: 武静, 西安交通大学管理学院博士研究生, 研究方向: 旅游需求预测、旅游智能推荐, E-mail: wujingsom@stu.xjtu.edu.cn; 赵二龙, 西安交通大学管理学院博士研究生, 研究方向: 旅游管理、数字营销, E-mail: Zhaoerlong@stu.xjtu.edu.cn; 通信作者: 孙少龙, 西安交通大学管理学院特聘研究员, 研究方向: 数字文旅、智慧运营, E-mail: sunshaolong@xjtu.edu.cn; 汪寿阳, 发展中国家科学院院士、国际系统与控制科学院院士, 中国科学院数学与系统科学研究院特聘研究员, 研究方向: 金融系统工程、经济分析与预测, E-mail: sywang@amss.ac.cn.

Multi-source Heterogeneous Data-driven Tourism Demand Forecasting Approach amid Post-COVID-19 Era

WU Jing¹, ZHAO Erlong¹, SUN Shaolong¹, WANG Shouyang^{2,3,4}

(1. School of Management, Xi'an Jiaotong University, Xi'an 710049, China; 2. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; 3. Center for Forecasting Science, Chinese Academy of Sciences, Beijing 100190, China; 4. School of Entrepreneurship and Management, ShanghaiTech University, Shanghai 201210, China)

Abstract Accurate tourism demand forecasting plays an important role in high quality development of tourism industry, especially short- and medium-term tourist arrivals forecasting is crucial for the tourism resource allocation and emergency management, the COVID-19 epidemic has led to structural changes in the tourism demand, posing new challenges to tourism demand forecasting and warning. This study focuses on whether fusing multi-source heterogeneous data can improve the performance of tourism demand forecasting approach amid post COVID-19 era, in order to explore this problem, firstly, a large number of variables affecting the fluctuation of tourism demand are extracted from online reviews generated by tourists, internet search data of tourists' attention and real-time online news data of tourist destinations, and topic model, sentiment analysis and feature engineering methods are adopted to deal with them in this study; secondly, mixed data sampling model with different data frequencies is used to fuse the variables to get multimodal fusion characteristics for forecasting; finally, fusing multi-source heterogeneous data-driven SARIMA-MIDAS approach is proposed for tourism demand forecasting. The main object of this study is tourist arrivals forecasting of Hong Kong amid post COVID-19 era. The empirical results reveal that fusing multi-source heterogeneous data-driven SARIMA-MIDAS approach can achieve the best forecasting performance amid post COVID-19 era. Therefore, the results of this study can provide a new solution for tourism demand forecasting and decision support for relevant government agencies and practitioners.

Keywords tourism demand forecasting; multi-source heterogeneous data; mixed-frequency data fusion; user-generated content; sentiment analysis

1 引言

随着我国经济的快速发展,我国旅游业规模也随之不断扩大,它在国民经济中的重要地位日益凸显.然而,新冠疫情的爆发对旅游业造成了巨大的冲击,这对旅游业的高质量发展带来了挑战(韦鸣秋等(2021)).因此,精准的旅游需求预测,不仅可以保证资源的高效配置和优质的服务,而且有利于快速调整旅游服务的供给,避免出现供求不平衡的现象.在后疫情时期,为了提高旅游需求预测的精度,本研究也对现存研究工作做了梳理,主要归纳为预测方法改进与预测影响因素探究(阮文奇等(2019),张嫚和黄凌云(2019),Yang et al.(2022),Xie et al.(2021)).

旅游需求的预测方法主要包括时间序列模型 (Li et al. (2018)), 计量经济模型 (Volchek et al. (2019), Zhang et al. (2021)), 机器学习技术 (Li et al. (2020), Bi et al. (2021), Höpken et al. (2021), Xie et al. (2021)). 但是, 新冠疫情的爆发导致旅游需求发生结构性变化, 这对旅游需求预测提出新挑战. 准确及时的旅游需求预测作为资源配置、应急管理、旅游政策设计和实施及旅游恢复判断的前提条件, 在 COVID-19 大流行期间发挥着重要作用 (刘培学等 (2022)). 因此, 后疫情时代旅游需求复苏预测已成为旅游管理研究重点关注的课题. 已经有不少学者来研究提高旅游需求的预测性能, 主要涉及到计量经济学和判断方法的组合 (Zhang et al. (2021)), 基于情景的判断方法 (Liu et al. (2021), Qiu et al. (2021), Kourentzes et al. (2021)), 基于专家判断的概率预测方法 (Athanasopoulos et al. (2022)) 和机器学习技术 (Li et al. (2022)). 表 1 对部分疫情期间旅游需求预测文献基本信息进行了归纳.

在预测影响因素的探究方面, 随着数字技术的快速发展, 旅游服务呈现出智能化和网络化的特征 (宋海岩和吴晨光 (2022)). 以往关于旅游需求预测的文献强调了预测影响因素在提高旅游需求预测准确性方面的重要性. 这些预测影响因素主要涉及经济相关变量、网络搜索数据、社交媒体帖子 (推特、在线评论、在线新闻、游客分享照片等) (于翠翠等 (2022), Li et al. (2020), Tian et al. (2021), Li et al. (2018), Wen et al. (2019), Bi et al. (2022), Hu et al. (2022), Volchek et al. (2019), Önder et al. (2019), Gunter et al. (2019)). 如 Sun et al. (2019) 利用旅游者关注的网络搜索数据结合历史游客量对北京的旅游需求进行预测. Yang et al. (2022) 利用旅游者关注的网络搜索数据对 74 个国家的游客量进行预测. Park et al. (2021) 采用旅游目的地实时在线新闻数据对中国香港的游客量进行预测. 然而, 很少有研究探究旅游者生成在线评论数据对旅游需求的影响 (Hu et al. (2022)). 在数字经济时代, 多源异构数据为预测方法提供了新的数据来源, 如何利用多源异构数据进行预测成为了一个亟待解决的问题 (Bi et al. (2020), 洪永淼和汪寿阳 (2021), 陈凯华 (2022), 孙少龙等 (2022), 邹松桦和刘秀丽 (2022)). 一方面, 这些数据由于数据结构和数据频度的不一致性往往给旅游需求预测带来一定的挑战, 例如, 旅游需求数据是月度或者季度数据, 而旅游者关注的网络搜索数据其反映游客对旅游目的地的关注程度是日度数据, 旅游者生成的在线评论数据揭示游客偏好、情绪以及目的地和景点的受欢迎程度和旅游目的地实时的在线新闻反映了旅游目的地的实时情况; 另一方面, 旅游需求数据和旅游者关注的网络搜索数据属于结构化的数据, 而在线评论数据和在线新闻数据属于非结构化文本数据, 不同类型数据的融合也增加了旅游需求预测的难度. 现有的研究方法大多将高频变量聚合为低频变量或将低频变量转化成高频变量, 这种同频化处理导致信息缺损或虚增、模型估计效率低下和存在偏差. 混频建模方法的优势在于把高频变量以及低频变量放在同一模型系统中进行统一建模分析, 同时考虑外生变量的混频特质, 混频建模方法以此减少高频数据转化为低频数据的信息损失问题, 大数据背景下数据的多源和复杂性使得混频建模方法成为有效建模工具之一 (秦梦和刘汉 (2019), Wen et al. (2021), Önder et al. (2019)). 特别是, 目前现有的研究很少采用融合多源异构数据结合混频方法进行旅游需求预测. 因此, 后疫情时期多源异构混频数据融合是否能够提升旅游需求预测方法的性能是一个值得探索的研究工作.

表 1 后疫情时期旅游需求预测研究工作基本信息归纳

研究者	目的地	数据频度	变量数据	预测方法	评价准则	研究背景
Zhang et al. (2021)	中国香港	季度	GDP, CPI, exchange rates, visitor arrivals atseven competing destinations	judgmental method	Breusch-Godfrey LM tests, Breusch-Pagan test, Jarque-Bera test, Ramsey RESET test, Co-integration test	16 个客源市场到 中国香港的游客量
Kourentzes et al. (2021)	20 个旅游目的地	季度	gross domestic product growth, PPP, implied PPP conversion rate	judgmental method	MASE, MAE	20 个目的地的国家国际游客量
Qiu et al. (2021)	20 个旅游目的地	季度	GDP, CPI, exchange rates	judgmental method	ASE	20 个旅游目的地地客流量
Liu et al. (2021)	20 个旅游目的地	季度	CORE index	judgmental method	MASE	20 个旅游目的地地客流量
Yang et al. (2022)	74 个旅游目的地国家	日度	Google trends, change in daily flight departures from international airport, policy, twitter data, daily number of confirmed COVID-19 cases	LASSO	MAE, RMSE, MAPE	74 个国家的每日旅游量
Bi et al. (2021)	九寨沟、四姑娘山	日度	tourist arrivals	CNN-LSTM	MAE, MAPE, RMSE	九寨沟和四姑娘山这两个中国著名景点的每日游客量
Li et al. (2022)	欧洲国家、北京	月度、日度	tourist arrivals	ST-FGCN	MAE, RMSE, MAPE, IR, DM	欧洲国家的游客量、北京景点小时游客量
Zhao et al. (2022)	四姑娘山	日度	tourist arrivals	TS-GBR, TS-SVR	RMSE, MAE, MAPE, DS, DM	四姑娘山每日游客量
Li et al. (2022)	夏威夷	月度	tourist arrivals	CEEMDAN-LSTM	MAE, RMSE, MAPE, WSR, SPA	夏威夷国内游客量
Li et al. (2023)	中国香港	月度	tourist arrivals, Baidu index and Google trends	SAE-Bi-GRU	MAE, SMAPE, RMSE	中国香港入境游客量

注: gross domestic product (GDP); consumer price index (CPI); purchasing power parity (PPP); COVID-19 risk exposure (CORE); least absolute shrinkage and selection operator (LASSO); convolutional neural network based on long short term memory neural network (CNN-LSTM); spatial-temporal integrated with graph convolutional network (ST-GCN); trajectory similarity integrated with gradient boosting regressor (TS-GBR); trajectory similarity integrated with support vector regression (TS-SVR); complete ensemble empirical mode decomposition with adaptive noise integrated with long short term memory neural network (CEEMDAN-LSTM); stacked auto-encoders integrated with bi-directional gated recurrent unit neural network (SAE-Bi-GRU); lagrange multiplier (LM); regression specification error test (RESET); mean absolute scaled error (MASE); mean absolute error (MAE); absolute scaled errors (ASE); root mean square error (RMSE); mean absolute percentage error (MAPE); improvement ratio (IR); Diebold-Mariano (DM); directional symmetry (DS); Wilcoxon signed-rank (WSR); superior predictive ability (SPA); symmetric mean absolute percentage error (SMAPE).

针对以上提到的研究问题和以往研究的空白,本研究通过融合旅游者关注的网络搜索数据、旅游目的地实时的在线新闻数据和旅游者生成的在线评论数据等不同频率的多源异构预测因子,提出了 SARIMA-MIDAS 方法用于后疫情时期中国内地到中国香港的游客量预测. 实验结果揭示出: 1) 旅游者关注的网络搜索数据、实时在线新闻数据和旅游者生成的在线评论数据都可以提高预测精度; 2) 同时融合旅游者关注的网络搜索数据、实时在线新闻数据和旅游者生成的在线评论数据能够显著提高预测精度; 3) 混频的多源异构数据融合的 SARIMA-MIDAS 预测方法在后疫情期间能够取得最佳的预测表现. 因此,该研究的结果可为旅游需求预测提供一种新的解决方案,为相关政府机构和从业者提供决策支持. 本研究工作的贡献主要有两个方面: 1) 对于旅游者生成的在线评论数据,利用 SnowNLP 库和 StanfordCoreNLP 库生成在线评论的情感极性得分; 2) 首次将新冠疫情期间在线评论的情感强度和评论量预测因子应用于旅游需求预测中.

本研究剩余部分安排如下: 第 2 章介绍了本研究用到的方法和预测框架; 第 3 章是实证研究和结果分析,最后,第 4 章给出结论与启示.

2 方法及预测框架

2.1 结构主题模型

主题模型 (topic modeling) 属于无监督的机器学习方法,是自然语言处理等领域用来发现一系列文档中抽象主题的一种统计模型. Blei et al. (2003) 提出的潜在狄利克雷分布 (latent Dirichlet allocation, LDA) 是一个包含文档-主题-词三个层次的贝叶斯模型, LDA 广泛应用于文本主题挖掘、文本分类聚类、文本检索、文本推荐等,目前是应用最广泛的主题模型之一. LDA 原理是基于文本主题之间两两不相关的特性,但是,在模型的拟合中仅依靠文本本身来挖掘语义主题,没有将文档已有的属性作为参数纳入,并且没有考虑到协变量与主题之间的关系. 为了解决 LDA 模型存在的问题,Roberts et al. (2014) 提出了结构化主题模型 (structural topic model, STM), STM 和 LDA 同属于生成模型的一种,与 LDA 不同的是 STM 模型在文档生成过程中,可以同时考虑多个文档协变量 (文档来源、文档发布时间等) 之间的联系来探究文档协变量与文档主题分布之间的隐含关系,在实际过程中,协变量对文档主题内容和主题流行度都有一定的影响. STM 模型的结构示意图如图 1 所示.

STM 模型的具体生成过程如下:

1. 基于文档协变量 X_r , 通过广义线性模型映射文档和主题之间的关系, 即,

$$\theta_r \sim \text{LogisticNormal}(\mu_r, \Sigma),$$

其中, $\mu_r = \Gamma'X_r'$, $\Gamma = (\gamma_1, \gamma_2, \dots, \gamma_{K-1})$ 是协变量对主题分布影响生成的回归系数矩阵, $\gamma_k \sim N(0, \sigma_k^2)$, Σ 是 $(K-1) \times (K-1)$ 的协变量矩阵.

2. 利用基准词分布 (m_v), 主题偏差 ($k_{r,v}^{\text{topic}}$), 协变量偏差 ($k_{yr,v}^{\text{cov}}$) 及其交互项偏差 ($k_{yr,k,v}^{\text{int}}$) 生成主题-词之间的分布, 即 $\beta_{rkv} \propto \exp(m_v + k_{r,v}^{\text{topic}} + k_{yr,v}^{\text{cov}} + k_{yr,k,v}^{\text{int}})$.

3. 生成文档中的各个词:

- 1) 从文档的主题分布中生成词的主题, 即 $Z_{r,n} \sim \text{Multinomial}(\theta_r)$;
- 2) 从主题的概率分布中生成词, 即 $W_{r,n} \sim \text{Multinomial}(\beta_{rk1}, \beta_{rk2}, \dots, \beta_{rkV})$.

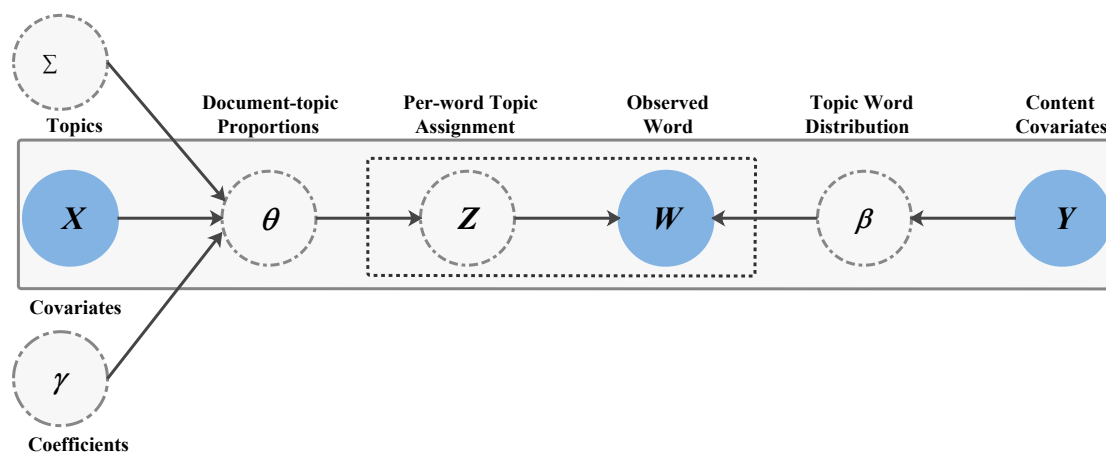


图 1 STM 的结构示意图

在使用 LDA 或 STM 模型识别文本中隐含的主题之前, 最重要的步骤是主题数量的选择, 因为主题数量是一个固定的参数, 其影响模型的计算结果. STM 主题模型的 SearchK 函数中能够综合 Held-Out (保留文档可能性)、Semantic Coherence (语义一致性)、residuals (残差)、Lower Bound (下界) 等指标来确定主题数量, 本研究利用 Held-Out 和 Semantic Coherence 来确定旅游目的地实时在线新闻的主题数量, 采用 R 软件中 stm Package 挖掘出旅游目的地实时在线新闻的主题分布.

2.2 广义动态因子模型

以往研究工作表明深度学习技术可以直接对高维变量进行建模预测, 但是对于计量经济模型, 其会引起模型的多重共线性和过拟合等问题. 为了解决这些问题, 在计量建模之前, 必须对高维变量进行降维, 这可使用数据压缩方法处理, 例如, 主成分分析方法、因子模型、动态因子模型等. Li et al. (2017), Li et al. (2020) 采用广义动态因子模型 (generalized dynamic factor model, GDFM) 对研究中的高维变量数据进行线性压缩使其产生复合指标, 实证结果也验证 GDFM 的优越性. 在本研究中, 将采用 GDFM 对旅游者关注度变量和在线实时新闻主题分布变量进行压缩降维. GDFM 是 Forniet 等人于 2000 年首次提出主要解决数据压缩当中公共因子的估计问题 (Forni et al. (2000)), 该方法既有动态因子模型的动态性, 也具有近似因子模型允许异质性部分相关的特性, 可以从大规模数据中提取共同的不可观测的因子. 令随机变量由 $\{X_{it}, i = 1, 2, \dots, n; t = 1, 2, \dots, T\}$ 表示, 其广义动态因子模型可以表示为:

$$X_{it} = \chi_{it} + \varepsilon_{it}, \quad (1)$$

$$\chi_{it} = b_{i1}(L)f_{1t} + b_{i2}(L)f_{2t} + \dots + b_{iq}(L)f_{qt}, \quad (2)$$

其中 χ_{it} 和 ε_{it} 分别表示公共部分和异质部分, 公共部分由公共因子 f_{jt} ($j = 1, 2, \dots, q$) 线性表示, 并且公共因子服从向量自回归, L 为滞后算子, q 表示公共因子数.

在该研究中, 采用 MATLAB 软件中 gdfm Package 对旅游者关注度变量和在线实时新闻话题分布变量进行压缩降维, 进而构建出游客关注度指数和新闻主题强度指数, 在进行

GDFM 之前, 需要对公共因子数进行确定, 本研究根据累积方差贡献率大于 90% 来确定因子个数.

2.3 SARIMA-MIDAS

旅游者关注的网络搜索数据、实时在线新闻数据和旅游者生成的在线评论数据的频率一般高于游客量数据. 当高频率、低频率变量同时存在时, 现有的大多数方法通常将高频变量聚合为低频变量或者把低频变量变成高频变量, 这同频化处理会导致信息缺损或虚增、模型估计效率低下和存在偏差. 为了解决此问题, Ghysels et al. (2004) 提出的混频抽样模型能解决变量不同频时的预测问题. MIDAS 在旅游研究中的应用仍处于起步阶段, 只有少数研究工作, 因此, 在本研究中, MIDAS 非常适合于月度游客量、日度搜索查询、周度在线新闻和周度在线评论的建模研究. MIDAS 基本形式如下:

$$Y_t = \beta_0 + \beta_1 \sum_{i=1}^m L_{\text{HF}}^i \omega(i; \Theta) X_t + \eta_t, \quad (3)$$

其中, Y_t 是游客量, L_{HF} 是高频项的滞后算子, $\omega(i; \Theta)$ 是一个多项式, 在滞后 i 时将权重分配给高频变量 X_t , m 是高频变量的最大滞后, η_t 是白噪声.

Wen et al. (2021) 提出 SARIMA-MIDAS 方法, 该方法认为 η_t 是 SARIMA 过程, 因此 SARIMA-MIDAS 方法基本形式如下:

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d\eta_t = \Theta(B^m)\theta(B)\varepsilon_t. \quad (4)$$

常见的权重函数有指数型 Almon 函数、Almon 多项式函数、Beta 形式、Step-function 形式等, 其中指数型 Almon 函数是目前最常用的权重, 在旅游需求预测中已经验证其具有一定的优越性. 因此, 在本研究中将采用指数型 Almon 函数为权重函数, 本研究采用 Python 中的 statsmodels 库实现该方法.

2.4 本研究流程框架

本小节主要介绍基于多源异构数据驱动的后疫情时期旅游需求预测方法研究的总体结构框架, 该框架如图 2 所示, 从图 2 可以看出, 基于多源异构数据驱动的后疫情时期旅游需求预测方法研究由四个步骤组成分别是数据收集、指标构建、模型构建和模型估计, 接下来简要介绍一下这四个步骤:

1) 数据收集. 数据收集是基于多源异构数据驱动的后疫情时期旅游需求预测方法研究的基础. 考虑到数据的可获得性、可用性和可靠性. 本研究收集到的数据主要包括结构化数据和非结构化数据, 其中结构化数据是旅游者关注的网络搜索数据, 非结构化数据主要是游客旅游过程中生成的在线文本评论数据 (酒店、景点) 和旅游目的地实时在线新闻数据. 在本研究中, 利用 Python 接口在百度指数网站上采集了 PC 端和手机端的旅游者关注的网络搜索数据, 同时使用 Python 在 China Daily 网站上爬取了旅游目的地实时在线新闻数据和在 TripAdvisor 网站爬取了中英文旅游者生成的在线评论数据.

2) 指标构建. 本研究对数据收集部分采集到的数据进行数据变换、数据融合等处理. 对于旅游者生成的在线评论数据, 本研究使用 StanfordCoreNLP 库来计算英文在线评论的情

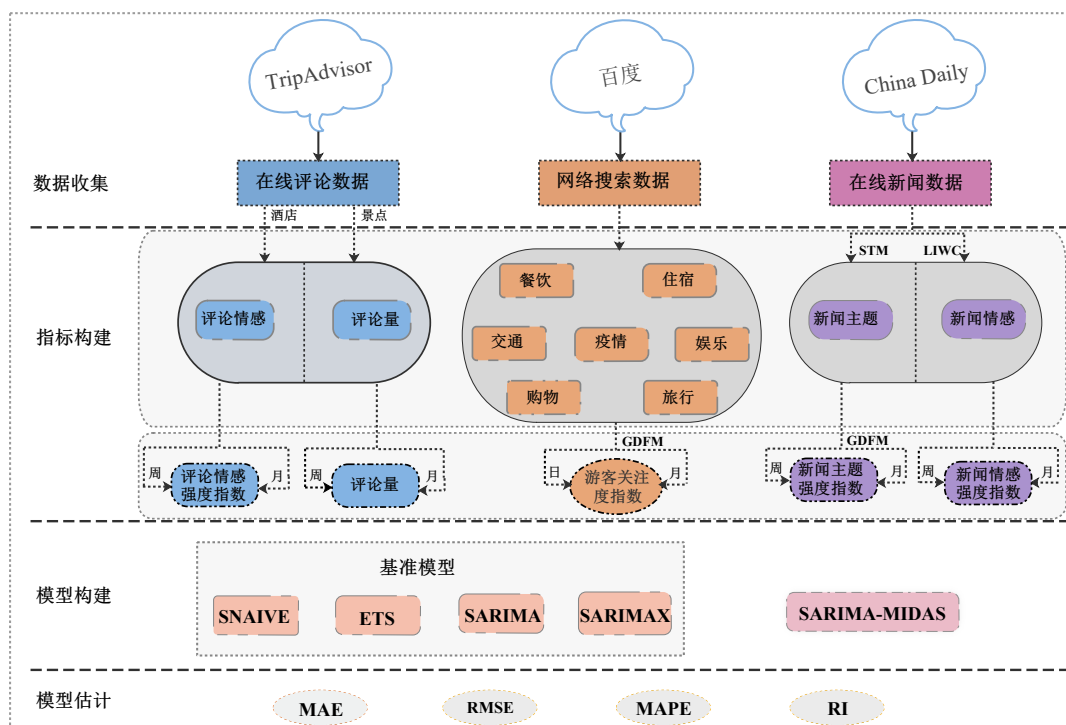


图2 基于多源异构数据驱动的后疫情时期旅游需求预测方法研究的总体框架

感极性分数,同时利用 SnowNLP 库来计算中文在线评论的情感极性分数.对于旅游目的地实时在线新闻数据,研究采用结构话题模型 (STM) 得到实时在线新闻的主题分布 (主题流行度),同时使用 LIWC 得到在线新闻的情感极性分数.为了得到周度和月度的在线评论情感极性分数、评论量、新闻的话题分布和新闻的情感分数,以及月度的旅游者关注的网络搜索数据,本研究对于月度的数据就是把当月的日度数据进行相加,对于周度的数据参考 (Hu et al. (2022)) 进行计算,通过计算得到了周度和月度的评论量和评论情感强度指数,以及新闻的情感强度指数.然而,由于周度和月度的话题分布以及日度和月度的旅游者关注的网络搜索数据都是以高维形式的数据表示,直接用于预测方法会产生过拟合或维数灾难的问题,所以本研究采用 GDFM 方法对新闻的主题分布和旅游者关注的网络搜索数据进行降维处理,得到周度和月度的新闻主题强度指数、日度和月度的旅客关注度指数.

3) 模型构建.经过指标构建之后,对构建的指标进行训练集和测试集的划分并根据训练集进行建模.研究以游客关注度指数、新闻主题强度指数、新闻情感强度指数、评论量和评论情感强度指数作为预测方法的自变量,旅游需求常以游客量为主要研究内容,所以本研究依然将游客量作为预测方法的因变量.本研究工作利用 SARIMA-MIDAS 来建立自变量和因变量之间的关系,从而得到旅游需求预测结果,同时把 SNAIVE、ETS、SARIMA 和 SARIMAX 作为基准方法.

4) 模型估计. 研究采用平均绝对误差 (MAE)、均方根误差 (RMSE) 和平均绝对百分比误差 (MAPE) 三个评价准则对不同预测方法的预测精度进行评估, MAE、RMSE 和 MAPE 值越小表示预测方法的预测性能越好. 同时为了直观地比较两个方法预测精度, 也使用改进程度 (RI) 指数来比较两个预测方法的相对提升性能.

3 实证检验

3.1 实验数据

中国香港是世界上著名的旅游目的地, 其文化多样性和购物机会每年吸引成千上万的游客, 尤其是来自中国内地的游客. 因此, 本研究将中国香港作为研究目标, 主要对中国内地到中国香港的游客量进行建模预测. 可以从 Wind Database (<http://www.wind.com.cn/>) 采集到中国内地到中国香港的月度游客量数据. 数据是从 2011 年 1 月到 2022 年 8 月. 从图 3 可以初步观察到游客量具有季节性和波动性, 特别是 2020 年 2 月出现了“断点”的现象.

本研究从百度指数 (<http://index.baidu.com/>) 收集了旅游者关注的日度网络搜索数据, 这些网络搜索数据的起止时间与游客量的时间一样. 关键词的选取将遵循以下三个阶段过程, 这样的选取准则可以有效地选择百度指数的关键词, 进而构建游客关注度指数.

1) 种子关键词选取. 依照旅游计划的考虑, 选取 26 个基准关键词, 其主要类别包括旅行、交通、住宿、购物、娱乐、餐饮和疫情七个方面. 表 2 列出了这 26 个基准关键词, 并给出了相应的类别.

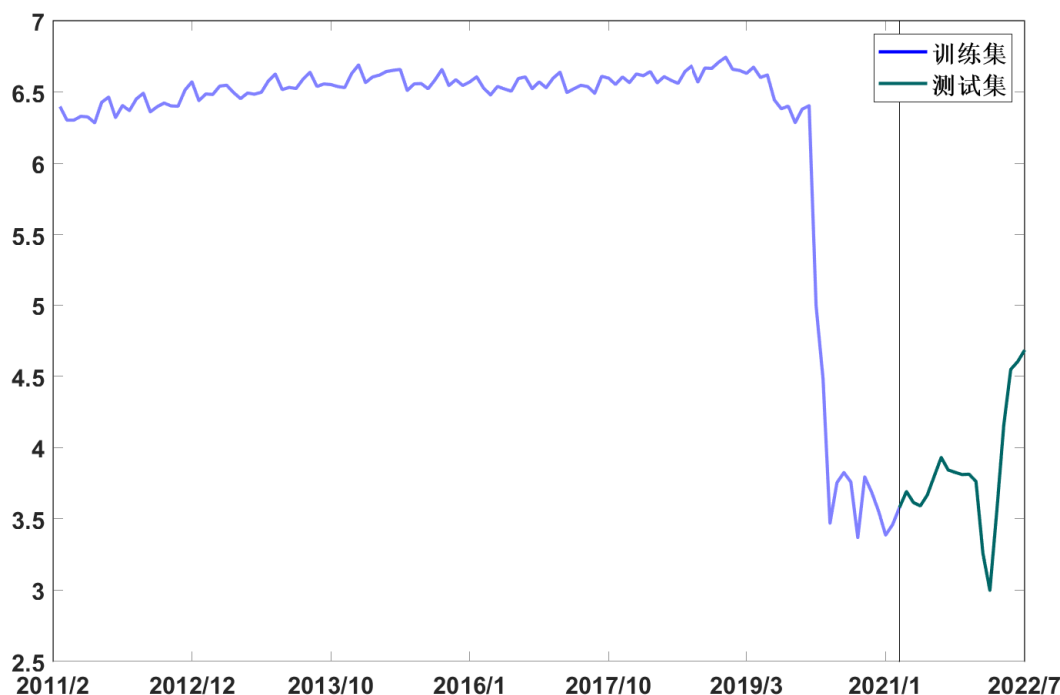


图 3 内地到中国香港每月的游客到达人数

2) 搜索引擎数据采集. 在百度指数中搜索了 26 个关键词作为种子关键词, 记录关键词所对应的相关搜索序列, 同时, 搜索引擎会根据其内部的算法给出与种子关键词相关联的推荐关键词; 然后, 迭代推荐关键词作为下一轮关键词, 重复进行多轮, 这个步骤一直迭代到关键词不可用或数据量极低为止, 最终得到百度索引关键词 89 个.

3) 旅游者关注的网络搜索数据预处理及指标的构建. 首先对 89 个旅游者关注的网络搜索数据进行按月求和得到月度的数据, 然后利用 GDFM 对旅游者关注的网络搜索数据进行压缩降维处理, 得到复合的指标, 即游客关注度指数. 因为旅游者关注的网络搜索数据是实时的数据, 为了后面的建模, 所以形成日度和月度的游客关注度指数.

对于旅游目的地实时在线新闻数据的选择, 本研究采用的在线新闻数据是从 China Daily (<http://www.chinadaily.com.cn/>) 通过 Python 爬虫采集到在线新闻数据. 这些新闻数据的起止时间与游客量的时间一样, 共收集到 48337 条数据. 得到新闻数据之后, 本研究使用 STM 模型从在线新闻文本中提取潜在主题, 并且利用 STM 模型的 SearchK 函数中的 Held-Out 和 Semantic Coherence 确定了在线新闻的主题数量为 80, 并得到了这些主题的主题分布, 主题主要包括经济、政治、社会问题、文化、国际事件、新冠疫情、行业问题、抗议问题等, 80 个主题分布也就是 80 个解释变量, 直接用于建模会引起过拟合等问题, 为了解决这一问题, 本研究利用 GDFM 进行变量压缩降维处理, 得到的复合指标是 80 个主题分布的加权. 对于新闻数据情感极性分数的计算, 本研究采用 LIWC 软件, 该软件是定量分析文本内容的词类常用的工具. 特别地, LIWC 可以计算出文本中不同类别词汇的使用比例, 如因果词、情感词、认知词等在整个文本中的使用比例. 因此, 在本研究中, 利用 LIWC 来计算在线新闻的情感极性分数, 由于实时新闻数据在解释旅游需求时不是连续或稳定的, 因此生成了周度和月度的新闻主题强度指数和新闻情感强度指数.

对于游客在旅游过程中生成的在线评论数据, 本研究是从 TripAdvisor (<https://www.>

表 2 关于中国香港旅游的百度搜索引擎种子关键词

序号	关键词	序号	关键词	序号	关键词
	旅行		交通		住宿
1	香港旅游	12	香港飞机场	19	香港住宿
2	香港天气	13	港澳通行证	20	香港酒店预订
3	香港旅游公司	14	香港地铁	21	香港酒店预订
4	香港旅游指南	15	香港地图	22	香港酒店
5	香港旅行社				
	娱乐		购物		餐饮
6	香港旅游景点	16	香港购物	23	香港美食
7	香港娱乐	17	香港购物攻略	24	香港小吃
8	香港海洋公园	18	香港购物清单	25	香港美食攻略
9	香港迪士尼乐园			26	香港美食推荐
	疫情				
10	COVID-19				
11	香港疫情				

tripadvisor.com) 通过 Python 爬虫采集到中文和英文的旅游者生成的在线评论数据. 主要是以景点和酒店的在线评论为数据源, 对于景点评论选择了评论数最多的 11 个景点, 分别是香港海洋公园、香港迪士尼乐园、香港维多利亚港、太平山山顶、山顶缆车、天坛大佛、香港天际线、赤柱市集、女人街、庙街夜市、香港电车, 共收集到 95766 条数据. 对于酒店评论选择了 5 星级、4 星级和 3 星级每级评论数最多的前五名, 共收集 15 个酒店的在线评论数据 59274 条. 本研究采用 StanfordCoreNLP 库来计算英文在线评论的情感极性分数, 同时利用 SnowNLP 库来计算中文在线评论的情感极性分数. 由于在线评论数据在解释旅游需求时不是连续或稳定的, 特别是在疫情期间, 因此生成了周度和月度的评论情感强度指数和评论量. 上述三种不同类型的数据统计情况如表 3.

表 3 多源数据基础信息的统计结果

数据类型	数据来源	内容	数据格式	数据量
旅游者生成的在线评论	https://www.tripadvisor.com/	景点与酒店评论	文本数据	155040 条
旅游者关注的网络搜索数据	http://index.baidu.com/	89 个关键词	结构化数据	411714 条
旅游目的地实时在线新闻数据	http://www.chinadaily.com.cn/	新闻文本	文本数据	48337 条

3.2 基准方法和预测评价准则

为了检验本研究提出的多源异构数据驱动的后疫情时期旅游需求预测方法预测性能, 选用常用的四种基准方法, 包括 SNAIVE, 指数平滑模型 (ETS), 季节性差分自回归滑动平均模型 (SARIMA) 和 SARIMAX.

为了评估和比较方法的预测精度, 采用文献中常见的三种评价准则, 包括平均绝对误差 (mean absolute error, MAE), 平均绝对百分比误差 (mean absolute percentage error, MAPE) 和均方根误差 (root mean square error, RMSE). 具体计算公式如下:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |x_i - \hat{x}_i|, \quad (5)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2}, \quad (6)$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|x_i - \hat{x}_i|}{x_i}, \quad (7)$$

其中, x_i 和 $\hat{x}_i$ 分别是 i 时刻游客量的实际值和预测值, n 表示测试集的数据量. 以上评价准则测量了预测方法的预测精度, 其值越小表示预测方法的预测性能越好. 此外, 为了直观地比较两个预测方法的精度, 本研究计算了 B 方法相比于 A 方法的改进程度 (relative improvement, RI), 计算公式如下:

$$\text{RI}_{\text{Model}_B}^{\text{Model}_A} = \frac{\text{MAPE}(\text{Model}_A) - \text{MAPE}(\text{Model}_B)}{\text{MAPE}(\text{Model}_A)}. \quad (8)$$

对于 RI 的结果, $\text{RI} > 0$, 说明 B 方法的预测效果相比于 A 方法有所改进, 而且值越大, 改进幅度越大; 如果 $\text{RI} < 0$, 说明 B 方法没有改进, 而且绝对值越大, B 方法预测效果越差.

3.3 实证结果

为了验证不同方法的性能, 该研究将数据划分为: 2011 年 1 月至 2021 年 2 月为训练数据, 2021 年 3 月至 2022 年 8 月为测试数据, 测试集主要包括疫情后期的时间段, 研究工作主要做旅游需求的提前 1 个月、2 个月和 3 个月的预测, 模型的设定如表 4 所示。

其中, Y_t 代表目前游客量; $VAI_{Monthly}$, VAI_{Daily} 分别代表月度和日度的游客关注度指数; $NTI_{Monthly}$, NTI_{Weekly} 分别代表月度和周度的新闻主题强度指数; $NSI_{Monthly}$, NSI_{Weekly} 分别代表月度和周度新闻情感强度指数; $RSI_{Monthly}$, RSI_{Weekly} 分别代表月度和周度在线评论情感强度指数; $RV_{Monthly}$, RV_{Weekly} 分别代表月度和周度在线评论量。为了验证游客关注度指数、新闻主题强度指数和新闻情感强度指数、评论量和评论情感强度指数对旅游需求预测方法的提升作用, 对 SARIMA 和 SARIMAX₁, SARIMAX₂, SARIMAX₃ 的结果进行对比分析。同样地为了验证 MIDAS 的优势本研究采用 SARIMAX_i 和 SARIMA-MIDAS_i ($i = 1, 2, 3$) 的结果进行对比。

根据以上实证部分的介绍, 表 5 至 7 分别是提前 1 个月、2 个月和 3 个月的预测比较结果, 具体如下。

从表 5 到 7 的结果可以明显看出, 本研究所提出多源异构数据驱动的后疫情时期旅游需求预测方法在不同的预测步长下的预测精度都大幅优于其他基准方法的表现。具体的结果及其分析如下: 从提前一个月的预测结果来看, 旅客关注度指数提高了旅游需求的预测精度, 且随着新闻主题强度指数和新闻情感强度指数, 以及评论量和评论情感强度指数的纳入, 预测效果都得到了提升, 同时预测结果也验证了 MIDAS 可以同时考虑了这些外生变量的混频特质, 以此减少高频数据转化为低频数据的信息损失问题进而提高预测的精度, 本研究所提出的 SARIMA-MIDAS₃ 方法的预测结果的 RMSE, MAE 和 MAPE 较其他基准方法有着不同程度的提高, 特别是较基准方法 SNAIVE, ETS 和 ARIMA 的预测结果具有明显的改进。同样的, 在步长为 2 和 3 的实证验证中同样存在以上的结论。为了进一步比较所提出方法的优越性, 分别计算 SARIMA-MIDAS₃ 与其他对照方法的 RI 提升比率, 结果如表 8 所示。

从表 8 的结果可以看出所提出的多模态驱动后疫情时期旅游预测方法在不同的预测步长上相对于其他基准方法在预测精度上都得到了提升, 特别对较传统的 SNAIVE 和 ETS 时间序列模型来说预测精度提高的程度有着明显的改进, 且 SARIMA-MIDAS_i ($i = 1, 2, 3$) 预测方法相对于其他方法具有明显的优势。在提前 1 个月的预测结果中本研究提出的预测方法较 SNAIVE 在 RMSE, MAE 和 MAPE 上分别提高了 75.1%, 73.9% 和 74.2%, 相对于 SARIMA-MIDAS₂ 在 RMSE, MAE 和 MAPE 上分别提高了 13.9%, 15.5% 和 13.5%。同样地在提前 2, 3 个月的预测中本研究提出的方法较其他方法有 8.0% 到 67.6% 的提升, 整体来说, 所提出的预测方法在所有对照方法中表现最佳。

4 结论与启示

旅游需求预测是旅游管理研究重要课题之一, 提高旅游需求预测性能是研究者和旅游从业者高度关注的问题。基于此, 本研究通过融合旅游者关注的网络搜索数据、旅游目的地实时的在线新闻数据和旅游者生成的在线评论数据等不同频率的多源异构预测因子 (游客关注度指数, 新闻主题强度指数和新闻情感强度指数, 及评论量和评论强度指数), 提出 SARIMA-

表 4 模型设定及形式

模型名称	模型形式
SNAIVE: 基于历史游客量信息的旅游需求预测	$\hat{Y}_t = Y_{t-m}$
ETS: 基于历史游客量信息的旅游需求预测	ETS(e, t, s)
SARIMA: 基于历史游客量信息的旅游需求预测	$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d Y_t = \Theta(B^m)\theta(B)\varepsilon_t$
SARIMAX ₁ : 基于历史游客量信息和月度的游客关注度指数的旅游需求预测	$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d Y_t = \alpha + \sum_{i=1}^m \beta_i \cdot VAI_{Monthly_{t-i}} + \Theta(B^m)\theta(B)\varepsilon_t$
SARIMAX ₂ : 基于历史游客量信息、月度的游客关注度指数、新闻主题强度指数和新闻情感强度指数的旅游需求预测	$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d Y_t = \alpha + \sum_{i=1}^m \beta_i \cdot VAI_{Monthly_{t-i}} + \sum_{i=1}^m \gamma_i \cdot NTI_{Monthly_{t-i}} + \sum_{i=1}^m \eta_i \cdot NSI_{Monthly_{t-i}} + \Theta(B^m)\theta(B)\varepsilon_t$
SARIMAX ₃ : 基于历史游客量信息、月度的游客关注度指数、新闻主题强度指数和新闻情感强度指数、以及评论量和评论情感强度指数的旅游需求预测	$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d Y_t = \alpha + \sum_{i=1}^m \beta_i \cdot VAI_{Monthly_{t-i}} + \sum_{i=1}^m \gamma_i \cdot NTI_{Monthly_{t-i}} + \sum_{i=1}^m \eta_i \cdot NSI_{Monthly_{t-i}} + \sum_{i=1}^m \lambda_i \cdot RV_{Monthly_{t-i}} + \sum_{i=1}^m \omega_i \cdot RSI_{Monthly_{t-i}} + \Theta(B^m)\theta(B)\varepsilon_t$
SARIMA-MIDAS ₁ : 基于月度的历史游客量信息和日度的游客关注度指数的旅游需求预测	$Y_t = \beta_0 + \beta_1 \sum_{d=1}^D \omega_1(d; \Theta) L_{HF}^d VAI_{Daily_t} + \eta_t, \quad \Phi(B^m)\phi(B)(1-B^m)^D \eta_t = \Theta(B^m)\theta(B)\varepsilon_t$
SARIMA-MIDAS ₂ : 基于月度的历史游客量信息、日度的游客关注度指数、周度的新闻主题强度指数和新闻情感强度指数的旅游需求预测	$Y_t = \beta_0 + \beta_1 \sum_{d=1}^D \omega_1(d; \Theta) L_{HF}^d VAI_{Daily_t} + \beta_2 \sum_{w=1}^W \omega_2(w; \Theta) L_{HF}^w NTI_{Weekly_t} + \beta_3 \sum_{w=1}^W \omega_3(w; \Theta) L_{HF}^w NSI_{Weekly_t} + \eta_t$ $\Phi(B^m)\phi(B)(1-B^m)^D \eta_t = \Theta(B^m)\theta(B)\varepsilon_t$
SARIMA-MIDAS ₃ : 基于月度的历史游客量信息、日度的游客关注度指数、周度的新闻主题强度指数和新闻情感强度指数、周度评论量和评论情感强度指数的旅游需求预测	$Y_t = \beta_0 + \beta_1 \sum_{d=1}^D \omega_1(d; \Theta) L_{HF}^d VAI_{Daily_t} + \beta_2 \sum_{w=1}^W \omega_2(w; \Theta) L_{HF}^w NTI_{Weekly_t} + \beta_3 \sum_{w=1}^W \omega_3(w; \Theta) L_{HF}^w NSI_{Weekly_t} + \beta_4 \sum_{w=1}^W \omega_4(w; \Theta) L_{HF}^w RV_{Weekly_t} + \beta_5 \sum_{w=1}^W \omega_5(w; \Theta) L_{HF}^w RSI_{Weekly_t} + \eta_t$ $\Phi(B^m)\phi(B)(1-B^m)^D \eta_t = \Theta(B^m)\theta(B)\varepsilon_t$

表 5 提前 1 个月预测结果对比分析				表 6 提前 2 个月预测结果对比分析			
方法	RMSE	MAE	MAPE	方法	RMSE	MAE	MAPE
SNAIVE	1.739	1.167	0.298	SNAIVE	1.731	1.164	0.297
ETS	1.147	0.916	0.225	ETS	1.270	1.024	0.252
SARIMA	0.797	0.608	0.147	SARIMA	0.819	0.694	0.171
SARIMAX ₁	0.661	0.522	0.128	SARIMAX ₁	0.755	0.599	0.146
SARIMAX ₂	0.629	0.497	0.122	SARIMAX ₂	0.717	0.570	0.139
SARIMAX ₃	0.544	0.421	0.102	SARIMAX ₃	0.639	0.526	0.129
SARIMA-MIDAS ₁	0.519	0.386	0.095	SARIMA-MIDAS ₁	0.612	0.497	0.122
SARIMA-MIDAS ₂	0.503	0.361	0.089	SARIMA-MIDAS ₂	0.609	0.489	0.120
SARIMA-MIDAS ₃	0.433	0.305	0.077	SARIMA-MIDAS ₃	0.560	0.440	0.108

表 7 提前 3 个月预测结果对比分析			
方法	RMSE	MAE	MAPE
SNAIVE	1.762	1.117	0.301
ETS	1.297	1.073	0.266
SARIMA	0.943	0.786	0.194
SARIMAX ₁	0.978	0.812	0.223
SARIMAX ₂	0.826	0.663	0.162
SARIMAX ₃	0.733	0.603	0.149
SARIMA-MIDAS ₁	0.713	0.600	0.148
SARIMA-MIDAS ₂	0.661	0.545	0.134
SARIMA-MIDAS ₃	0.595	0.479	0.118

表 8 SARIMA-MIDAS ₃ 和其他对照方法的 RI 结果分析								
方法	SNAIVE	ETS	SARIMA	SARIMAX ₁	SARIMAX ₂	SARIMAX ₃	SARIMA-MIDAS ₁	SARIMA-MIDAS ₂
提前 1 步预测对比								
RMSE	0.751	0.622	0.457	0.345	0.312	0.204	0.166	0.139
MAE	0.739	0.667	0.498	0.416	0.386	0.276	0.210	0.155
MAPE	0.742	0.658	0.476	0.398	0.369	0.245	0.189	0.135
提前 2 步预测对比								
RMSE	0.676	0.559	0.316	0.258	0.219	0.124	0.085	0.080
MAE	0.622	0.570	0.366	0.265	0.228	0.163	0.115	0.100
MAPE	0.636	0.571	0.368	0.260	0.223	0.163	0.115	0.100
提前 3 步预测对比								
RMSE	0.662	0.541	0.369	0.392	0.280	0.188	0.165	0.100
MAE	0.571	0.554	0.391	0.410	0.278	0.206	0.202	0.121
MAPE	0.608	0.556	0.392	0.471	0.272	0.208	0.203	0.119

MIDAS 方法用于后疫情时期中国内地到中国香港的游客量预测, 用 MAE, RMSE 和 MAPE 作为方法性能的评估指标, 预测的主要结果如下: 1) 旅游者关注的网络搜索数据、实时在线新闻数据和旅游者生成的在线评论数据都可以提高预测精度; 2) 将旅游者关注的网络搜索数据、实时在线新闻数据和旅游者生成的在线评论数据融合进预测方法能够显著提高预测精

度; 3) 混频的多源异构数据融合的 SARIMA-MIDAS 预测方法在后疫情期间能够取得良好的预测表现. 实证的结果表明, 本研究提出的基于不同频度的多源异构数据融合驱动的后疫情时期旅游需求预测方法能够提高预测精度, 有助于未来的研究通过更多不同模态的数据来提高预测能力.

本研究所提出的多源异构数据驱动后疫情时期旅游需求预测方法在理论和实践中具有一定的意义. 在理论层面, 由于疫情对旅游需求预测的影响, 造成常用方法预测结果的不稳定, 该研究基于文本智能支撑的多源异构数据驱动的方法可以为后疫情时期的预测方法提供一种范式; 在应用层面, 本研究的预测结果可以为中国香港旅游相关部门的政策制定提供重要的信息支持, 以便在后疫情时代增加当地旅游业的竞争力.

本研究所提出的预测方法除可以应用于旅游需求预测中, 也可以应用于其他领域, 如可以结合其他多源异构的数据对股票价格, 外汇汇率, 石油价格等预测预警. 当然本研究也存在一定的局限性, 如该工作只考虑中国香港的预测研究, 有待在其他旅游目的地对所提的预测方法进行鲁棒性验证, 同时该研究所选的非结构化数据类型主要是文本数据, 在以后研究中可以对其他非结构化数据 (图片、音频、短视频等) 进行探究.

参 考 文 献

- 陈凯华, (2022). 创新计量学: 理论与方法 [J]. 计量经济学报, 2(2): 209–227.
- Chen K H, (2022). Innovametrics: Theory and Methods[J]. China Journal of Econometrics, 2(2): 209–227.
- 洪永淼, 汪寿阳, (2021). 大数据、机器学习与统计学: 挑战与机遇 [J]. 计量经济学报, 1(1): 17–35.
- Hong Y M, Wang S Y, (2021). Big Data, Machine Learning and Statistics: Challenges and Opportunities[J]. China Journal of Econometrics, 1(1): 17–35.
- 邹松桦, 刘秀丽, (2022). 一种改进的 CEEMDAN-LSTM-Neural Prophet Net 模型: 用于 COVID-19 背景下我国月度消费预测 [J]. 计量经济学报, 2(3): 620–643.
- Huan S H, Liu X L, (2022). An Improved CEEMDAN-LSTM-Neural Prophet Net Model for Monthly China's Total Retail Sales of Consumer Goods Prediction under the Background of COVID-19[J]. China Journal of Econometrics, 2(3): 620–643.
- 刘培学, 王欢颖, 陈伟, 张建新, 刘泽华, (2022). 新冠疫情对旅游景区客源市场需求的空间分异影响 —— 以南京夫子庙景区为例 [J]. 地理科学, 42(7): 1250–1259.
- Liu P X, Wang H Y, Chen W, Zhang J X, Liu Z H, (2022). Spatially Divergent Impact of the COVID-19 Epidemic on Source Markets of Tourism Destination: A Case Study of the Confucius Temple in Nanjing[J]. Scientia Geographica Sinica, 42(7): 1250–1259.
- 秦梦, 刘汉, (2019). 百度指数、混频模型与三亚旅游需求 [J]. 旅游学刊, 34(10): 116–126.
- Qin M, Liu H, (2019). Baidu Index, Mixedfrequency Model and Sanya Tourism Demand[J]. Tourism Tribune, 34(10): 116–126.
- 阮文奇, 张舒宁, 李勇泉, 郑向敏, (2019). 中国赴泰旅游需求时空分异及其影响因素 [J]. 旅游学刊, 34(5): 76–89.
- Ruan W Q, Zhang S N, Li Y Q, Zheng X M, (2019). Spatiotemporal Differentiation and Influencing Factors of Chinese's Tourism Demand to Thailand[J]. Tourism Tribune, 34(5): 76–89.
- 宋海岩, 吴晨光, (2022). 新一轮科技革命与旅游需求分析和预测创新: 理论探讨与实践前沿 [J]. 旅游学刊, 37(10): 1–3.

- Song H Y, Wu C G, (2022). The New Technological Revolution and Innovation in Tourism Demand Analysis and Forecasting: A Theoretical Discussion and Practical Frontier[J]. Tourism Tribune, 37(10): 1-3.
- 孙少龙, 魏云捷, 黎建强, (2022). 基于在线外汇新闻情感挖掘的汇率预测研究 [J]. 计量经济学报, 2(2): 441-464.
- Sun S L, Wei Y J, Li J Q, (2022). Exchange Rate Forecasting with Online Forex News Sentiment Mining[J]. China Journal of Econometrics, 2(2): 441-464.
- 韦鸣秋, 白长虹, 张彤, (2021). 旅游目的地精益服务供给中的组织关系演进逻辑 —— 基于重庆、西安、杭州的跨案例比较研究 [J]. 管理世界, 37(7): 119-129.
- Wei M Q, Bai C H, Zhang T, (2021). The Evolution Logic of Organizational Relationship in Fine Service Supply of the Tourism Destination: A Cross-case Study Based on Chongqing, Xi'an, and Hangzhou[J]. Management World, 37(7): 119-129.
- 于翠翠, 寇红红, 孙少龙, (2022). 基于多语言搜索引擎数据的旅游需求预测研究 [J]. 系统科学与数学, 42(9): 2383-2398.
- Yu C C, Kou H H, Sun S L, (2022). Forecasting Tourism Demand with Multi-language Search Engine Data[J]. Journal of Systems Science and Mathematical Sciences, 42(9): 2383-2398.
- 张嫒, 黄凌云, (2021). 负面网络关注度对旅游业发展的影响——基于旅游需求的空间关联分析 [J]. 旅游学刊, 36(7): 81-91.
- Zhang M, Huang L Y, (2021). Tourism Development and Public Attention to Negative Online Information: Based on the Spatial Correlations of Tourism Demand[J]. Tourism Tribune, 36(7): 81-91.
- Athanasopoulos G, Hyndman R J, Kourentzes N, O'Hara-Wild M, (2022). Probabilistic Forecasts Using Expert Judgment: The Road to Recovery from COVID-19[J]. Journal of Travel Research. <https://doi.org/10.1177/00472875211059240>.
- Bi J W, Li C, Xu H, Li H, (2022). Forecasting Daily Tourism Demand for Tourist Attractions with Big Data: An Ensemble Deep Learning Method[J]. Journal of Travel Research; 61(8):1719-1737.
- Bi J W, Li H, Fan Z P, (2021). Tourism Demand Forecasting with Time Series Imaging: A Deep Learning Model[J]. Annals of tourism Research, 90: 103255.
- Bi J W, Liu Y, Li H, (2020). Daily Tourism Volume Forecasting for Tourist Attractions[J]. Annals of Tourism Research, 83: 102923.
- Blei D M, Ng A Y, Jordan M I, (2003). Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 3(Jan): 993-1022.
- Forni M, Hallin M, Lippi M, Reichlin L, (2000). The Generalized Dynamic-Factor Model: Identification and Estimation[J]. Review of Economics and Statistics, 82(4): 540-554.
- Höpken W, Eberle T, Fuchs M, Lexhagen M, (2021). Improving Tourist Arrival Prediction: A Big Data and Artificial Neural Network Approach[J]. Journal of Travel Research, 60(5): 998-1017.
- Hu M M, Li H Y, Song H Y, Li X, Law R, (2022). Tourism Demand Forecasting Using Tourist-generated Online Review Data[J]. Tourism Management, 90: 104490.
- Ghysels E, Valkanov R. (2004). The MIDAS Touch: Mixed Data Sampling Regression Models[R]. CIRANO Working Papers, 5(1), 512-517.
- Gunter U, Önder I, Gindl S, (2019). Exploring the Predictive Ability of LIKES of Posts on the Facebook Pages of Four Major City DMOs in Austria[J]. Tourism Economics, 25(3): 375-401.
- Kourentzes N, Saayman A, Jean-Pierre P, Provenzano D, Sahli M, et al. (2021). Visitor Arrivals Forecasts Amid COVID-19: A Perspective from the Africa Team[J]. Annals of Tourism Research, 88: 103197.
- Li C, Zheng W M, Ge P, (2022). Tourism Demand Forecasting with Spatiotemporal Features[J]. Annals

- of Tourism Research, 94: 103384.
- Li H Y, Hu M M, Li G, (2020). Forecasting Tourism Demand with Multisource Big Data[J]. Annals of Tourism Research, 83: 102912.
- Li M C, Zhang C Y, Wang S Y, Sun S L, (2022). Multi-scale Analysis-driven Tourism Forecasting: Insights from the Peri-COVID-19[J]. Current Issues in Tourism: 1–24.
<https://doi.org/10.1080/13683500.2022.2144151>.
- Li M C, Zhang C Y, Sun S L, Wang S Y, (2023). A Novel Deep Learning Approach for Tourism Volume Forecasting with Tourist Search Data[J]. International Journal of Tourism Research, 25: 183–197.
- Li S W, Chen T, Wang L, Ming C H, (2018). Effective Tourist Volume Forecasting Supported by PCA and Improved BPNN Using Baidu Index[J]. Tourism Management, 68: 116–126.
- Li X, Pan B, Law R, Huang X K, (2017). Forecasting Tourism Demand with Composite Search Index[J]. Tourism Management, 59: 57–66.
- Liu A, Vici L, Ramos V, Giannoni S, Blake A, (2021). Visitor Arrivals Forecasts Amid COVID-19: A Perspective from the Europe Team[J]. Annals of Tourism Research, 88: 103182.
- Önder I, Gunter U, Scharl A, (2019). Forecasting Tourist Arrivals with the Help of Web Sentiment: A Mixed-frequency Modeling Approach for Big Data[J]. Tourism Analysis, 24(4): 437–452.
- Park E, Park J, Hu M M, (2021). Tourism Demand Forecasting with Online News Data Mining[J]. Annals of Tourism Research, 90: 103273.
- Qiu R T R, Wu D C, Dropsy V, Petit S, Pratt S, et al. (2021). Visitor Arrivals Forecasts Amid COVID-19: A Perspective from the Asia and Pacific Team[J]. Annals of Tourism Research, 88: 103155.
- Roberts M E, Stewart B M, Tingley D, Lucas C, Leder-Luis J, et al. (2014). Structural Topic Models for Open-ended Survey Responses[J]. American Journal of Political Science, 58(4): 1064–1082.
- Sun S L, Wei Y J, Tsui K L, Wang S Y, (2019). Forecasting Tourist Arrivals with Machine Learning and Internet Search Index[J]. Tourism Management, 70: 1–10.
- Tian F, Yang Y, Mao Z X, Tang W Y, (2021). Forecasting Daily Attraction Demand Using Big Data from Search Engines and Social Media[J]. International Journal of Contemporary Hospitality Management, 33(6): 1950–1976.
- Volchek K, Liu A, Song H Y, Buhalis D, (2019). Forecasting Tourist Arrivals at Attractions: Search Engine Empowered Methodologies[J]. Tourism Economics, 25(3): 425–447.
- Wen L, Liu C, Song H Y, (2019). Forecasting Tourism Demand Using Search Query Data: A Hybrid Modelling Approach[J]. Tourism Economics, 25(3): 309–329.
- Wen L, Liu C, Song H Y, Liu H, (2021). Forecasting Tourism Demand with an Improved Mixed Data Sampling Model[J]. Journal of Travel Research, 60(2): 336–353.
- Xie G, Qian Y, Wang S Y, (2021). Forecasting Chinese Cruise Tourism Demand with Big Data: An Optimized Machine Learning Approach[J]. Tourism Management, 82: 104208.
- Yang Y, Fan Y W, Jiang L, Liu X H, (2022). Search Query and Tourism Forecasting During the Pandemic: When and Where Can Digital Footprints be Helpful as Predictors?[J]. Annals of Tourism Research, 93: 103365.
- Zhao E L, Du P, Sun S L, (2022). Historical Pattern Recognition with Trajectory Similarity for Daily Tourist Arrivals Forecasting[J]. Expert Systems with Applications, 203: 117427.
- Zhang H Y, Song H Y, Wen L, Liu C, (2021). Forecasting Tourism Recovery Amid COVID-19[J]. Annals of Tourism Research, 87: 103149.