基于随机森林算法的宏观经济先行指数体系构建

赵云飞　娄　峰　程　远*

内容摘要：本文选取 2001 年 1 月至 2022 年 6 月的数据作为样本，构建了国内生产总值（GDP）、价格指数和出口的宏观经济先行指数，并用 2022 年 7 月至 12 月的数据进行样本外预测。并将本文构建的随机森林（RF）指数体系与现存的指数进行效果对比，结果表明:（1）相比传统先行指数，RF 先行指数可以更好地反映宏观经济变量的变动趋势，对历史期的经济发展拐点和波动幅度均高度拟合。（2）RF 指数构建方法对于不同领先期数的设定和不同被预测变量的选择在样本内是稳健的。（3）RF 指数对名义 GDP、工业生产者出厂价格指数（PPI）和出口指数的样本外数据成功预测了走势和振幅，对于实际 GDP 和居民消费价格指数（CPI）的走势预测效果较好，但振幅的预测效果一般。综上，RF 先行指数体系和方法对现有指数的计算方法、操作流程的简便性、期数选择和变量选择的稳健性以及预测效果等多个方面都具有显著的改善。

关键词：随机森林模型；经济先行指数；经济预测；景气指数

中图分类号：F2　**文献标识码：**A　**文章编号：**1004-7794(2024)04-0003-13

DOI: 10.13778/j.cnki.11-3705/c.2024.04.001

一、引言与文献综述

准确预测未来经济运行情况对于企业生产经营和政府实施经济政策具有重要作用。因此，各国研究部门和商业机构均在积极研发可以预测宏观经济变量的模型和方法。其中，构建先行指数体系是普遍应用的判断经济运行、周期变化和预测拐点的主要方法之一。1917 年，哈佛大学利用股票市场、商品市场和金融市场共 17 个指标构建了能预测经济走势的“哈佛指数”，成功预测出未来几年的经济增长、衰退和复苏，但由于未能成功预测 1927—1933 年的经济危机而招致质疑。而后，美国国家经济研究局（NBER）于 1937 年根据时差变动关系从 500 个相关经济变量中筛选出 21 个指标，并运用扩散指数法编制了宏观经济指数。美国国家经济研究局的 Moore（1961）[1]划分了先行、一致和滞后指标，并基于扩散指数法编制了景气指数，在预测经济运行拐点方面有较好的效果，但是无法预测具体波动幅度。为了弥补该项不足，Shiskin et al.（1968）[2]提出了合成指数法，既能预测经济运行趋势，还可以预测波动幅度大小。经济合作与发展组织（OECD）[3]在筛选构建指数的指标过程中，兼顾经济学和统计学，并且尽可能地扩大指标覆盖的经济领域，于 1981 年开始构建各国经济合成先行指标 OECD 指数并至今保持每月更新。Stock

* 赵云飞，中国社会科学院大学应用经济学院博士研究生，研究方向为宏观经济与政策模拟。娄峰，2005 年毕业于对外经济贸易大学，获经济学博士学位，现为中国社会科学院数量经济与技术经济研究所教授、博士生导师，研究方向为宏观经济预测和政策模拟分析。程远（通讯作者），2017 年毕业于南开大学，获经济学博士学位，现为中国社会科学院数量经济与技术经济研究所副研究员，研究方向为宏观经济学，邮箱：chy805@126.com。本研究得到自然科学基金青年项目“资金流量视角下的金融与实体经济关系研究——资金可计算一般均衡模型构建及政策分析”（72204263）、“中国社会科学院经济大数据与政策评估实验室项目”（2024SYZH004）的资助。

et al.（1989）[4]综合多个经济领域的宏观经济时间序列并计算加权平均值，完善了美国国家经济研究局一致经济指数，而后利用 1959—1987 年的数据验证了该景气循环方法可以综合准确的反映经济增长情况。

国内相关研究始于 1980 年代。董文泉等（1987）[5]通过匹配波峰和波谷，构建了先行、一致和迟行指标，并论证了我国经济景气循环的存在性和拐点发生的大致时间，开创了我国经济循环测定、分析和预测问题的研究，但其多个指标之间缺乏联系，并且存在经济波动幅度仍无法准确预测的问题。陈磊等（1993）[6]在此基础上，利用主成分分析方法对指标做降维处理，并采用简单加权的方法编制了先行指数和一致指数，预测效果有了明显的提升，但指标库仍以工业以及工业品消费为主。而后，陈磊等（1994）[7]借鉴 Stock-Watson 景气指数，将指标库扩展到金融市场、劳动市场和商品市场等多个经济领域。随着经济发展和数据的完善，除 GDP 之外，价格水平、进出口贸易等多个宏观经济变量都纳入了预测目标当中。并且由于我国各地区、各行业的差异，分地区和分行业的先行指数构建也成为了重点关注的领域。

从具体做法来看，现存构建先行指数的过程大都借鉴 OECD 指数。OECD 指数先用线性插值法填充了缺失值，而后对指标的季节性趋势用 X12 方法做了调整，并且通过两次 HP 滤波处理依次过滤掉数据的长期趋势和高频噪音，最后对所有指标进行标准化处理，采用简单平均的方式合成 OECD 指数。之后，美国企业联合会（Conference Board）[8]构建的领先、一致和滞后指数在选取指标的标准和过程与 OECD 指数大体类似，而在指数构建过程中，CB 指数对于采用了自回归（AR）模型拟合值填充缺失值，并计算指标月度环比，以标准差的倒数为权重加权求和，目的是赋予波动较小的指标更大的权重。而后国内外应用较多的筛选指标的方法有峰谷对应、时差相关分析、K-L 信息量等。王恩德等（2008）[9]选择 K-L 信息量、时差相关分析、峰谷对应等方法，构建了辽宁省宏观经济景气分析系统。孙延芳等（2015）[10]通过类似步骤计算扩散指数预测经济运行拐点，计算合成指数预测波动幅度，构建了领先 4 个季度的建筑行业的景气指数。李学文等（2015）[11]在时差相关分析和峰谷对应法的基础上建立了回归分析模型，计算出指标的最佳滞后期数，对湖南省经济景气预警指数做出了研究。何建文（2018）[12]以 K-L 信息量为主要筛选方法，并利用主成分分析法获得各指标权重，最后利用加权平均得到的景气指数进行环比变化率的计算，构建了高新技术制造业景气指数。黄文静等（2021）[13]构建了中金领先指数，同样从经济理论和拐点匹配的角度对指标进行筛选，并借鉴了 OECD 的两次 HP 滤波处理步骤，采用简单平均法分别构建了领先 1 期、2 期和 3 期的领先指数，来量化中国经济复苏路径。徐寅等（2021）[14]利用信息比率阈值筛掉波动过于剧烈的指标，这一点与 CB 指数相似，即偏好于波动较小的变量，再结合跨期相关性和 K-L 散度选择指标领先期，参考 Moore 合成指数方法，对指标进行标准化后做简单平均再计算环比变化率，并将指标按照所在经济部门构建了综合、金融环境和实体 3 个先行指数。在指标筛选后，合成指数的方法也存在不同，如果没有特殊要求，一般采用标准化后简单平均的方式，也可以根据各指标重要性选择加权平均的方式得到指数初始值，而后将基准期指数值设为 100，通过计算标准化平均变化率不断迭代得到指数序列。如刘玉娇等（2020）[15]、狄浩林等（2022）[16]基于电力消费数据构建经济指数，在采用 K-L 信息量和跨期相关系数作为筛选指标依据的基础上，对指标计算标准化平均变化率后合成指数。

上述文献所使用的方法尚存在一定的改进空间。第一，各类方法对于构建经济先行指数所需经济变量的选择缺乏客观、一致的标准。各类文献在选取构建指数的经济变量时，会参考相关性、K-L 信息量、拐点匹配度等不同统计指标，并结合经济理论进行选择，不同的统计指标和标准分析评判经济变量得到的结论是不一致的甚至是矛盾的，导致所选取的经济变量组合存在较大的主观性。第二，在数据预处理过程中，以 OECD 指数为代表的多数研究都依赖于 HP 滤波来消除数

据难以拟合的高频波动，以提升预测效果。但用滤波处理数据无论在理论还是应用上都存在一些问题，从而影响到指数的计算：首先，对原始数据滤波提高了指标的平滑性，便于提升指数拟合的效果，但也导致原始数据包含的一些信息在这一过程中受到了损失；其次，宏观经济变量大多具有单位根，而 HP 滤波对非平稳数据倾向于放大经济周期，同时降低短期和长期的波动[17]；再次，对不同时间长度的数据进行滤波时，样本区间改变会导致相同数据的取值发生改变，尤其在样本两端尤为明显[18]；最后，HP 滤波的平滑参数的选取尚存在争议，惯例是根据序列的时间频率选择，但这种方法过于依赖于主观经验的判断，缺乏科学合理的设定标准。这些问题都影响到计算得到的指数的客观性和准确度。第三，对于构建的先行指数相对于实际经济指标的领先期，使用跨期相关性、拐点匹配法、K-L 信息量法等不同方法测得的领先期是相互矛盾的，同样缺乏一致的、普遍认可的确定指数领先期的标准。另外，一些指数合成方法将不同领先期的经济变量加总时会导致其波峰和波谷的不同步而互相抵消，反而掩盖了各自的波动趋势。

本文采用随机森林方法构建宏观经济先行指标，从而针对以上问题做出改进。随机森林方法是将多个决策树模型集成后对样本进行训练和测试的一种机器学习方法，由于其较少依赖经济学理论及经济学模型方程形式的设定，善于捕捉数据本身之间复杂多变的非线性数量关系[19-21]，从而广泛应用于经济数量分析。相比于现有构建经济先行指数的方法，利用随机森林方法有以下几个方面的优势：首先，随机森林方法中的特征值重要性反映了特征值对于目标变量的解释力，可以作为选择数据指标的客观标准；其次，随机森林方法不须对数据滤波就可以构建出较好反映宏观经济变量的变动趋势指数，规避了 HP 滤波等方法存在的问题，也更充分利用了原始数据包含的信息；最后，随机森林方法并非选取具有预测性和不同周期性的经济变量并将其进行简单的线性合成，而是寻找和拟合待估变量和指定领先期的目标变量之间的非线性数量关系，从而避免了测定领先期及合成不同周期经济变量时产生的问题。

在使用随机森林方法进行宏观经济预测方法方面，Alessi et al.（2011）[22]认为随机森林可以提供经济体的早期衰退预警，Biau et al.（2010）[23]用随机森林模型预测了欧洲 GDP，初步证实了随机森林在大型数据集预测中的作用。但是以上文献所使用的方法还无法直接用于构建经济先行指数。首先，构建经济指数需要在大量经济变量中筛选出对目标经济指标具有良好预测性的变量，以上文献仍然使用传统构建经济先行指数的方法选择变量，并未提供和随机森林的非线性分析方法相匹配的变量筛选方法。其次，构建经济先行指数要求确定甚至有目的地选择构建指数的领先期，这是仅对经济指标进行预测的现有研究无法提供的。本文将随机森林方法应用于构建经济先行指数，在随机森林的方法框架内提供了变量筛选、指数拟合以及预测效果评价等相契合的分析方法，对现有分析方法进行了拓展。

在宏观经济变量中，本文着重关注经济增长、价格水平和国际贸易三个方面以构建宏观经济先行指数体系。首先，GDP 是衡量一个经济体运转状况和刻画经济周期最全面的基础指标，包含了对经济周期测度的重要数据信息[24]，由于名义 GDP 衡量本期经济活动的综合结果，包括了物质产量因素和物价因素，而实际 GDP 排除了价格波动的影响，因此本文选择名义 GDP 和实际 GDP 作为衡量经济运行情况的变量。其次，价格水平的变动是宏观经济运行情况的重要衡量指标之一，其中，居民消费价格指数（CPI）衡量了与居民生活紧密相关的消费品价格和劳动要素价格水平，生产者出厂价格指数（PPI）衡量了工业企业产品出厂价格变动趋势和变动程度。并且考虑到自 2016 年至今，我国 CPI 和 PPI 出现过多次背离现象[25]，因此本文分别选择工业生产者出厂价格指数（PPI）和居民消费价格指数（CPI）同时作为衡量价格水平的变量。第三，自从改革开放以来，我国出口成为了拉动我国经济增长的重要引擎之一，因此本文关注了贸易领域中出口先行指数的构建。需要说明的是，虽然进口是国际贸易中一个重要的经济变量，但考虑到进口

更多内生于我国收入水平，其循环周期和国内经济周期较为一致，因此本文暂不对进口先行指数的构建进行讨论。

本文的创新点主要体现在以下两个方面。首先，本文将随机森林方法应用于宏观经济先行指数的构建，提供了在随机森林的模型框架内包括变量筛选、数据处理，指数构建以及预测效果评价等一系列具体方法，从而对现有构建经济先行指数的方法进行了改进。第二，本文利用随机森林方法计算了包含名义 GDP、实际 GDP、CPI、PPI 和出口等变量在内的先行指数指标体系，并构建了领先期分别为 3 个月、6 个月、9 个月的宏观经济先行指数，并对这一方法构建先行指数的预测效果及对不同目标变量和领先期的适用性进行了验证。

二、方法介绍

（一）决策树

决策树算法是一种数据分析中常见的机器学习方法，其目的是通过对现有数据的分析，通过递归过程训练出一棵泛化能力较强的树模型。由于本文选择的指标均为连续变量，因此应采用 C4.5 决策树算法，即二分法来分割数据[26]。在数据集中，每个特征值的每个可能取值都是一个数据节点，通过数据节点可以将原数据集分割为 2 个子数据集，在这个过程中，选择能使分割过程熵增最大的节点作为决策树的第一层节点。之后，按照同样的方法将子数据集继续分割，形成一个多层的决策树模型。

假设特征值数据集中含有 N 个特征值，总期数为 T，即：

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{TN} \end{bmatrix} \tag{1}$$

用数据集的信息熵来衡量样本集合 X 的混乱程度：

$$Entropy(X)=-\sum_{i=1}^{N} p_i \times \log p_i \tag{2}$$

其中，p_i 代表数据集中第 i 类的概率。

数据集中共有 $N\times T$ 个特征值的取值。每个取值都可以作为 1 个数据节点 x_{tn}，数据节点可以将总体数据集 X 分割为 2 个子数据集。即：

$$X_1=\{x_i \mid x_{ti} \leqslant x_{tn}, t=1,\cdots,T, i=1,\cdots,N\} \tag{3}$$

$$X_2=\{x_i \mid x_{ti} > x_{tn}, t=1,\cdots,T, i=1,\cdots,N\} \tag{4}$$

此时，称 x_{tn} 为决策树的叶子节点。于是问题就变为，选择一个最科学的叶子节点将样本分成两部分。希望通过样本的切分，尽可能地将相同特征的个体分到同一个小样本中，换句话说，就是希望分割出来的两个小样本内部混乱程度越低越好。用信息增益（Information Gain）来计算分割前后数据集混乱程度的差异。一般来说，信息增益越大，说明分割过程对于数据集的纯度提升越大[27]。

$$Gain(X,x_n)=\max_{x_{tn}}\left\{Entropy(X)-\sum_{\lambda\in\{1,2\}}\frac{\left|X_t^{\lambda}\right|}{|X|}Entropy(X_t)\right\} \tag{5}$$

其中，x_n 为选择的切割变量，选择 x_n 的某取值 x_{tn} 作为节点分割数据集 X 为 X_1 和 X_2，使得 X

的信息熵与 X_1，X_2 的信息熵和的差值最大。通过遍历所有数据节点，找到信息增益值最大的 x_{tn} 作为决策树的根节点，在分割后的子样本中重复以上步骤，来构造决策树模型。

（二）随机森林

在实际的回归和分类问题中，单一模型往往会受到参数选择、样本数据的局限性等问题的影响导致模型效果不好，集成算法就是通过多种算法和策略对同一个数据集进行回归或分类之后，取简单平均值或加权平均值作为结果，弥补了单一模型的局限性和偶然性。集成算法分为 3 种：并行集成 bagging 算法、串行集成 boosting 算法和 stacking 算法。本文所用的随机森林是一种决策树的并行集成 bagging 算法[28]。其中，“随机”二字有两层含义，第一是对训练样本进行 m 次有放回随机抽样形成采样集，再重复这个过程 k 次，形成 k 个包含 m 个样本的采样集。再从每个小采样集中随机抽取一定比例的指标来构建独立的决策树模型[29]，得到结果 $DT(y_k)$，样本和特征值的双重随机性使个体决策树之间的差异度进一步增加，最终提升算法的泛化性能。由于同一森林中不同的决策树之间性能相差很小，参考 Xu et al.[30]，将这种双重随机抽样法得到的决策树结果 $DT(y_k)$ 放到一起取简单平均值得到集成模型的结果 $RF(y)$：

$$RF(y)=\frac{1}{k}\sum_{i=1}^{k}DT(y_k) \tag{6}$$

三、数据处理

（一）被预测变量的处理

由于官方发布的名义和实际 GDP 数据均为季度值数据，本文先对 GDP 数据作季度同比，而后对 GDP 季度同比数据进行线性插值处理，生成名义 GDP 月度同比和实际 GDP 月度同比，以保证数据高频化后仍维持原有的走势；CPI 和 PPI 数据选择当月同比价格指数，无需再做处理；出口金额是当月值数据，将出口金额做当月同比处理。因此，本文分别对名义 GDP 月度同比、实际 GDP 月度同比、居民消费价格指数、工业生产者出厂价格指数和出口金额月度同比共 5 个变量构建随机森林模型。

（二）特征值

参考 OECD 指数的方法，选择的指标应在样本区间内无长时间缺失数据，且指标统计口径保持一致。考虑到统计频率越高、样本数量越多，模型效果越好，因此本文以月度数据为主，并将获取的季度数据通过线性插值法填充为月度数据。本文在 Choice 数据库的中国宏观数据和行业经济数据中，获取 364 条和经济相关的指标构建指标库并进行初步筛选。首先，以全国为视角，因此筛掉分行业数据和分地区数据；其次，由于年度数据统计跨度太长，且样本较少，不适合对月度数据做预测，本文也不采用；再次，后续预测需要获得指标的未来取值，因此删掉已经停止更新的数据；最后，对于同一数据的不同统计口径只选一条，选择优先级为月度同比>当月值>季度同比>季度值。初步筛选之后，获得相关日度数据 7 条，月度数据 191 条，季度数据 49 条，共计 247 条数据。选择的经济指标数量如下表 1。

而后，将日度数据取当月平均值、季度数据做线性插值处理变频为月度数据。本文参考兴业指数，将指标的月度数据转换为月度同比，并不再对同比数据进行季节调整。具体做法为：若原数据为同比或指数，则不作任何处理；若原数据为当月值，则计算月度同比；若原数据为月度累计值，则先做差分后计算月度同比。之后再次进行数据筛选，删除如下类型数据：（1）开始统计日期过晚，无法与待估变量做匹配；（2）存在跨度较大的缺值区间，线性插值无法准确拟合指标走势；（3）用其他方法合成的我国或其他国家的经济预测指数，如 OECD 先行指标、银行业景气

指数等等，只保留未经合成的一手数据。

筛选过后，剩余 7 条日度转月度数据、139 条月度数据和 38 条季度转月度数据共计 184 条数据作为备选特征值库，见表 2。

表 1　　变量选取　　（个）

经济领域	季度数据	月度数据	日度数据	经济领域	季度数据	月度数据	日度数据
工业	4	25		金融	8	35	
价格指数		5		就业与工资	1	1	
固定资产投资		14		景气指数	22	7	
对外经济贸易	1	4		人民生活水平	3	1	
汇率		4		社会科教		1	
利率	1			特色宏观数据		33	7
财政	6	19		行业经济数据	2	9	
国内贸易	1	15		共计	49	191	7
证券市场		18					

资料来源：东方财富 Choice 数据库。

表 2　　指标描述

变量	时间跨度	匹配特征值个数	样本数量
名义 GDP 月度同比	2001 年 1 月至 2022 年 6 月	94	258
实际 GDP 月度同比	2008 年 3 月至 2022 年 6 月	136	178
居民消费价格指数	2001 年 1 月至 2022 年 6 月	94	258
工业生产者出厂价格指数	2001 年 1 月至 2022 年 6 月	94	258
出口金额月度同比	2014 年 1 月至 2022 年 6 月	172	102

首先，将备选特征值库的 184 条数据作为特征值，将 5 个待估变量提前 6 期作为被解释变量，分别对 5 个待估变量进行建模；而后，随机抽取 75%的样本作为训练集得到决策树模型，剩余 25%的样本作为测试集，用来计算指标和作图，评价模型的优劣。重复多次上述过程，将多个决策树集成为随机森林，并取平均值作为回归结果。通过观察，每个待估变量的匹配特征值过多，特征值重要性<0.01 的变量对模型的贡献可以忽略不计，因此需要计算特征值重要性并排序，对于每个变量只保留特征值重要性≥0.01 的指标作为解释变量，这个过程为模型精简过程，为了验证精简的有效性，本文选择测试集的 MAPE 作为评价指标用来对比精简前后的模型。表 3 汇报了通过提取特征值重要性≥0.01 的指标后，占样本个数 25%的测试集的 MAPE 对比，可以看出，精简后所有模型的 MAPE 都降低了，说明精简过程可以提高模型效果。而后将每个待估变量选择的解释变量中删除多余的相似指标，比如“名义汇率”和“实际汇率”如果都被选为解释变量，则保留特征值重要性较大的变量，为了防止过拟合问题，对决策树进行剪枝，重新构建随机森林模型并生成拟合值。

表 3　　变量筛选前后模型效果对比

变量	指标个数	精简前 MAPE	精简后 MAPE
名义 GDP 月度同比	11	1.02%	0.86%
实际 GDP 月度同比	13	19.08%	12.46%
居民消费价格指数	12	30.00%	23.38%
工业生产者出厂价格指数	14	108.19%	74.45%
出口金额月度同比	12	445.05%	387.32%

四、模型结果及分析

（一）构成指标选取

表 4 汇报了各模型选择的特征值，可以看出特征值涵盖了我国工业、零售消费品、货币政策、汇率、期货和股指等多个经济领域，部分模型还引入了其他国家的进出口金额和失业率等经济变量作为特征值。

表 4　　模型指标

名义 GDP	实际 GDP	CPI
工业增加值	不纳入广义货币的存款	基础货币余额
工业企业利润总额	国债交易成交金额	实收资本
美国季调失业率	饮料类商品零售值	流通中现金（M0）
名义有效汇率指数	布伦特原油期货结算价	美国工业生产指数
广义货币供应量（M2）	房地产开发投资完成额	金融机构各项贷款余额
欧盟季调失业率	工业企业利润总额	伦敦金属交易所铅期货收盘价
伦敦金属交易所锌期货收盘价	欧元区 CPI	欧盟失业率
房地产开发企业到位资金	原油产量	日本出口额
美国出口额	日本 CPI	全国商品零售价格指数
外汇储备	非税财政预算收入	企业商品交易价格指数
日本进口额	基础货币余额	城镇居民人均可支配收入
	粮油食品类商品零售值	美国商品调查局现货指数
	外汇储备	
PPI	出口金额	
工业企业利润总额	银行间市场人民币交易成交金额	
企业商品交易价格指数	金融机构外汇各项存款余额	
广义货币供应量乘数	工业企业利润总额	
深证成分指数	外商直接投资	
美国出口额	利润总额	
农村居民人均可支配收入	城镇房屋竣工面积	
狭义货币（M1）	采购经理指数	
伦敦金属交易所锌期货收盘价	其他类金融机构总资产	
实际有效汇率指数	企业商品交易价格指数	
发电量	房地产开发新增固定资产投资	
出口金额	全国期货市场成交金额	
日本 CPI	不纳入广义货币的存款	
上证综合指数		
境内上市公司流通市值/广义货币 M2		

（二）预测效果分析

1. 模型评价指标。

为了多方面评价模型优劣，参考蒋锋等[31]，本文选择可决系数（R^2）、平均绝对误差（MAE）和均方误差（MSE）作为衡量模型效果的标准，具体指标含义如下，其中 y_t 为真实值，$\bar{y}$ 为样本均值，$\hat{y}_t$ 为模型拟合值。

可决系数（R^2）是指已经解释的变异在其总变异中所占的比率：

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y})^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \quad (7)$$

平均绝对误差（*MAE*）是指预测值和观测值之间绝对误差的平均值：

$$MAE=\frac{1}{n}\sum_{t=1}^{n}\left|\left(y_t-\hat{y}_t\right)\right| \tag{8}$$

均方误差（*MSE*）是指预测值和观测值之间差异的样本方差：

$$MSE=\frac{1}{n}\sum_{t=1}^{n}\left(y_t-\hat{y}_t\right)^2 \tag{9}$$

模型的评价指标和测试集、训练集的样本数如表 5。

表 5　　模型效果检验指标

变量名	R^2	*MSE*	*MAE*	测试集样本数	训练集样本数
名义 GDP 月度同比	90.52%	0.0061	0.0039	64	189
实际 GDP 月度同比	76.35%	0.0193	0.0117	42	125
消费品价格指数	89.70%	0.2315	0.1734	64	189
工业生产者出厂价格指数	87.29%	0.5421	0.3936	64	189
出口金额月度同比	57.48%	0.0358	0.0243	25	72

2. 领先期为 6 个月的指数预测情况分析。

图 1 汇报了样本区间内各模型先行指数和变量真实值的走势对比图。从图 1 中可以看出，在国际金融危机和新冠肺炎疫情期间，GDP 的同比增速在样本区间内出现过 2 次大的增速下滑。受美国次贷危机影响，名义 GDP 同比在 2008 年下降，实际 GDP 则在 2010 年出现第一个波峰。第 2 次波动在 2020 年初新冠肺炎疫情暴发之后，我国经济遭受比次贷危机更强的负面冲击，名义 GDP 和实际 GDP 月度同比一度下降至负值，此后在 2020 年 6 月恢复为正增长。价格指数方面，CPI 和 PPI 在 2012 年之前走势基本相同，波峰和波谷出现的位置也大致相近，到了 2012 年之后 PPI 迅速下降至负值，并于 2015 下半年开始反弹，直至 2017 年初出现波峰，而 CPI 则在 2012 年之后一直保持小幅度波动的低通胀走势。出口金额方面，同比数据在 2020 年之前波动幅度较大，在 2014 年 8 月、2015 年 9 月和 2017 年 8 月达到极大值点，而在疫情期间同比数据波动幅度减小，呈现出稳定的增长趋势。

综合来看各模型的走势对比图可以发现本方法合成的指数具有以下特征。第一，在 GDP、价格指数和出口同比的模型中，当变量出现较为明显和持续的增长或下降时，指数可以很好地预测到各变量的拐点出现时间和方向。第二，在 GDP 和价格指数模型中，当同比数据在小区间内出现多次剧烈波动时，先行指数也可以提前捕捉波动拐点。第三，本模型合成的先行指数更多反映的是目标指标的趋势性特征，因此先行指数在波峰的最高点往往低于实际数据的相应最高点，而先行指数在波谷的最低点往往高于实际数据相应位置的最低点。例如，2020 年之后新冠肺炎疫情期间名义 GDP 和实际 GDP 的情况，以及 CPI 在 2004 年和 2008 年左右的波峰的情况。第四，在一些实际数据较为平稳的期间，先行指数可能会出现小范围波动。例如，实际 GDP 同比在 2011—2012 年和 2016—2017 年两个时期的情况，这可能是由于合成指数的多个宏观变量具有不同的变动趋势所造成的。

3. 不同领先期先行指数的预测情况比较。

为了验证随机森林算法对于构建不同领先期的先行指数效果的稳健性，本文选择相同的指标筛选和模型拟合的方法，构建了名义 GDP 同比增长率的领先 3 个月、6 个月和 9 个月的先行指数进行比较。从图 2 可以看出，不同领先期数的先行指数均捕捉到了目标经济指标变动的主要趋势，具有较好的预测效果。同时，先行指数相比较目标变量的领先期也较为准确，符合设定的情况。

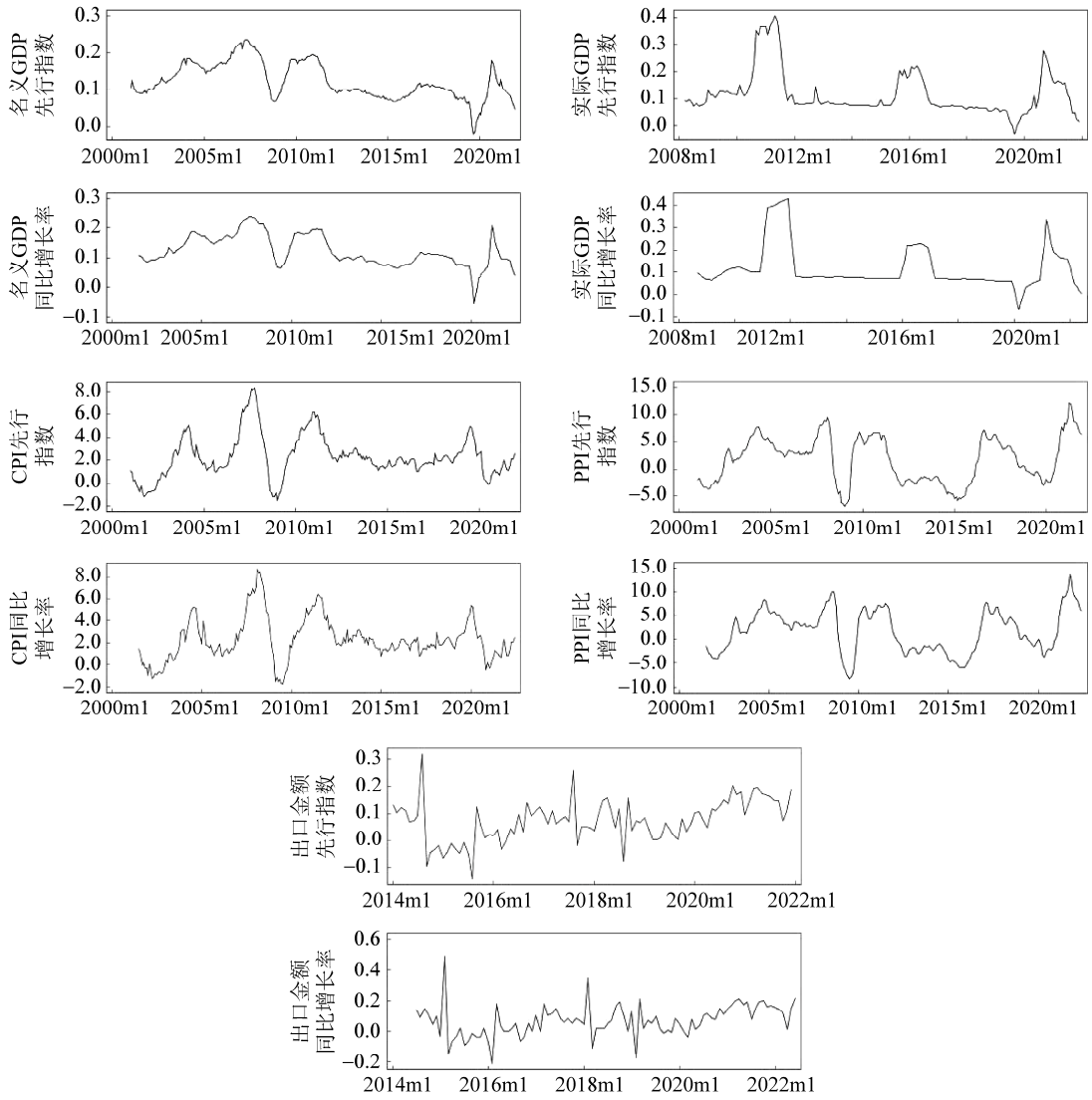

图 1　模型样本内先行指数预测效果

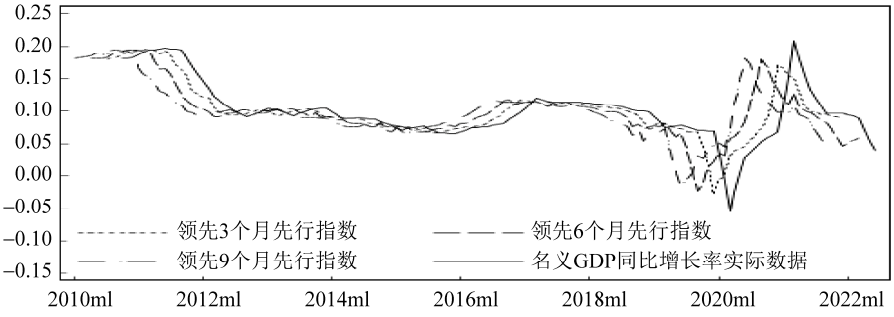

图 2　名义 GDP 不同领先期先行指数预测效果

4. 与其他建模方式的预测情况比较。

为了对比随机森林算法和现有方法的差异，本文以构建名义 GDP 先行指数为例，主要参考 Shiskin et al.（1968）以及徐寅等（2021），以拐点匹配为主，综合利用跨期相关性、K-L 信息量

等方法并结合经济理论选取经济指标合成名义 GDP 同比增长率的先行指数。选取的经济指标包括 6 个金融领域指标和 14 个实体经济领域指标共 20 个经济指标，如表 7 所示。

表 6　　名义 GDP 不同领先期先行指数指标选取

领先期 3 个月指标选取	领先期 9 个月指标选取	领先期 3 个月指标选取	领先期 9 个月指标选取
工业增加值	实际有效汇率指数	基础货币余额	外汇储备
美国出口额	广义货币供应量（M2）	烟酒类商品零售值	城镇固定资产投资累计值
城镇居民人均消费性支出	美国失业率	企业商品交易价格指数	存款性公司货币
人民币名义有效汇率指数	欧元区消费者信心指数	发电量	伦敦金属交易所锌期货收盘价
日本季调失业率	工业增加值	伦敦金属交易所锌期货收盘价	固定资产投资完成额
城镇固定资产投资	人民币各项贷款余额	实收资本	房地产开发企业到位资金

表 7　　指标选取

一级分类	二级分类	变量含义	一级分类	二级分类	变量含义
金融指标	信用	狭义货币（M1） 广义货币供应量（M2） 金融机构企业存款余额 金融机构各项贷款余额	实体经济指标	投资	固定资产投资合计资金来源 固定资产投资新建资金来源
	资产价格	上证综合指数 上证所股票成交金额		工业产量	钢材产量 水泥产量 发电量
实体经济指标	投资	房地产开发资金来源 商品房销售面积 商品房销售额 固定资产投资完成额		贸易	沿海主要港口货物吞吐量 民航货邮运输量
				海外景气度	美国对中国出口金额 欧元区制造业 PMI 美国费城联储制造业指数

合成指数方法参考 Shiskin et al.[2]。首先，对单个指标 $X_i(t)$ 进行 HP 滤波，并计算指标历史差分的绝对值均值 $Y_i(t)$：

$$Y_i(t)=\frac{\sum_{i}^{n}\left|X_i(t)-X_i(t-1)\right|}{n} \tag{10}$$

接下来，计算指标的标准化差分 $Z_i(t)$：

$$Z_i(t)=\frac{\left|X_i(t)-X_i(t-1)\right|}{Y_i(t)} \tag{11}$$

最后，对指标进行等权求和得到 $V(t)$，令初值 $I(0)$为 100，计算环比变化率合成指数 $I(t)$：

$$I(t)=I(t-1)\times\frac{200+V(t)}{200-V(t)} \tag{12}$$

利用传统方法合成的经济指数和随机森林方法合成指数的情况效果如图 3 所示。可以看出，两种方法均可捕捉到 GDP 同比增长率的主要波动，并且在走势上和目标序列基本保持一致。相比而言，随机森林方法构建的经济指数更加具有优势，主要体现在以下三个方面。第一，随机森林方法对经济波动和经济走势的预测更加精准。传统方式构建的指数虽然可以捕捉大幅度的波动的长期经济发展趋势，但对中小幅度的波动和经济较为平稳时期的经济走势的判断不够准确。第二，相比于传统方式，随机森林方法构建的指数对经济波动的具体幅度也有较好的反映。第三，随机森林方法构建的指数预测不同时期经济波动的领先期更加均匀，基本等于设定的领先期。传统方式构建的指数预测不同经济波动的领先期并不一致，有些经济波动很早便会提示，有些波动则较晚才会反映。

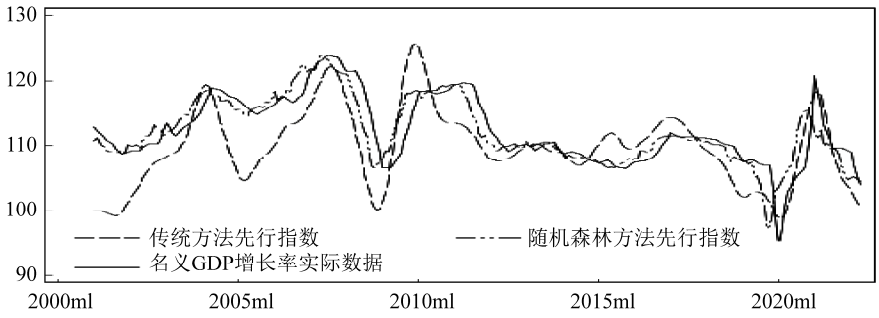

图 3　GDP 指数预测效果图

（三）样本外预测结果

图 4 汇报了样本外预测结果，为了更好地识别先行指数的效果，图中加入了 2015 年 1 月至 2022 年 6 月的样本内数据，其中先行指数区间为 2022 年 1 月至 2022 年 6 月，即图中竖线右侧的先行指数为样本外预测。从图 4 可以看出，各变量先行指数从 2022 年 1 月至 6 月的总体走势和 2022 年 7 月至 12 月的实际增长率数据的走势大致吻合，说明本文的方法具有较好的预测效果。

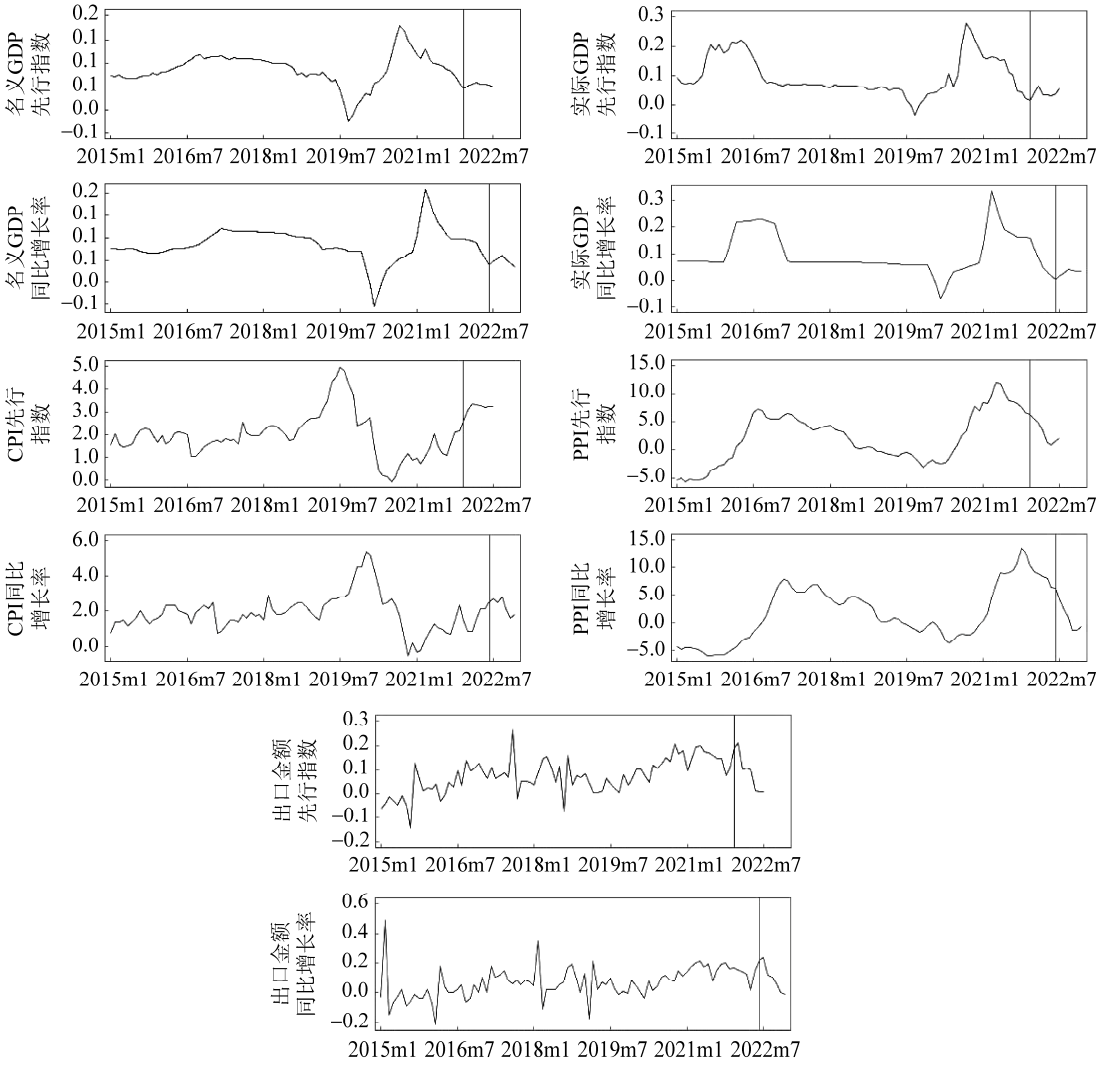

图 4　模型样本外先行指数预测效果

五、总　结

本文选取我国 2001 年 1 月至 2022 年 6 月的数据作为样本，运用随机森林算法构建了 GDP、价格指数和出口的领先期为 6 个月的先行指数，并用 2022 年 7 月至 12 月的数据进行样本外预测。此外，本文还比较了不同领先期的经济指数及与其他方式构建的经济指数预测效果的差异。结果表明：（1）利用随机森林方法合成的 RF 宏观经济先行指数在样本区间内可以较好地反映宏观经济变量的变动趋势，对经济发展拐点和波动幅度的拟合效果良好，并且样本内的拟合效果对于不同经济变量的选择是稳健的。（2）分别以 3 个月、6 个月、9 个月为领先期的 3 种 RF 指数的领先效果及预测能力都很好，说明在不同领先期长度的研究需求下，RF 先行指数的领先效果也是稳健的。（3）在样本外区间，RF 指数对名义 GDP、工业生产者出厂价格指数（PPI）和出口指数预测了变量走势方向和波动幅度，对于实际 GDP 和居民消费价格指数（CPI）预测了变量走势方向，但波动幅度的预测不太准确，因此 RF 指数构建方法在未来仍有改进的空间。

根据本文的分析可以发现，采用随机森林模型构建的经济先行指数克服了现有方法的诸多不足，对经济具有较好的预测性。此外，本文的结果也显示了，将机器学习中随机森林模型等非线性数量分析模型用于构建经济先行指数具有较好的效果。本文的结论可以为后续进一步研究提供参考。

参考文献

[1] Moore G H. Business Cycle Indicators[M]. Princeton, NJ: Princeton University Press, 1961.

[2] Shiskin J, Moore G H. Composite Indexes of Leading, Coinciding, and Lagging Indicators, 1948–67[M]. Supplement to NBER Report One. NBER, 1968: 1-8.

[3] Stock J H, Watson M W. New Indexes of Coincident and Leading Economic Indicators[J]. NBER Macroeconomics Annual, 1989(4): 351-394.

[4] 董文泉, 郭庭选, 高铁梅. 我国经济循环的测定、分析和预测(I)——经济循环的存在和测定[J]. 吉林大学社会科学学报, 1987(3): 1-8.

[5] 陈磊, 吴桂珍, 高铁梅. 主成分分析与景气波动——对 1993 年我国经济发展趋势的预测[J]. 数量经济技术经济研究, 1993(7): 33-37.

[6] 陈磊, 高铁梅. 利用 Stock-Watson 型景气指数对宏观经济形势的分析和预测[J]. 数量经济技术经济研究, 1994(5): 53-59.

[7] Organization for Economic Co-operation and Development. OECD System of Composite Leading Indicators[R]. 2021.

[8] Conference Board Inc. Handbook on Cyclical Composite Indicators[R]. 2017.

[9] 王恩德, 陈飞, 梁云芳. 辽宁省宏观经济景气分析系统的研究与应用[J]. 统计与决策, 2008(9): 27-30.

[10] 孙延芳, 胡振. 中国建筑业景气指数的合成与预测[J]. 统计与决策, 2015(11): 40-42.

[11] 李学文, 李明贤. 湖南省宏观经济景气预警指数的构建[J]. 系统工程, 2015, 33(12): 101-106.

[12] 何健文. 中国高技术制造业创新景气指数研究——基于主成分法[J]. 科技管理研究, 2018, 38(11): 6-12.

[13] 黄文静, 张文朗, 周彭, 等. 复苏斜率将如何演变？——详解中金经济领先指数[R]. 中金公司研究部, 2021.

[14] 徐寅, 宫民. 如何构建中国经济先行指数[R]. 兴业证券, 2021.

[15] 刘玉娇, 宋坤煌, 王向. 基于电力大数据的经济景气指数分析[J]. 电信科学, 2020, 36(6): 166-171.

[16] 狄浩林, 于海洋, 宋玉鑫, 等. 基于用电数据的经济景气指数模型及实证[J]. 统计与决策, 2022, 38(22): 20-24.

[17] 梁琪, 滕建州. 中国经济周期波动的经验分析[J]. 世界经济, 2007(2): 3-12.

[18] Hamilton J D. Why you Should Never Use the Hodrick-Prescott Filter. Review of Economics and Statistics, 2018, 100(5): 831-843.

[19] Boulesteix, Anne-Laure; Janitza, Silke; Kruppa, Jochen; König, Inke R. Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. Wiley Interdisciplinary Reviews: Data

Mining and Knowledge Discovery, 2012, 2(6): 493-507.
[20] Gawthorpe K. Random Forest as a Model for Czech Forecasting[J]. Prague Economic Papers, 2021, 30(3): 336-357.
[21] Nicolas Woloszko, 2020. “Adaptive Trees: a New Approach to Economic Forecasting,” OECD Economics Department Working Papers 1593, OECD Publishing.
[22] Alessi L, Detken C. Quasi Real Time Early Warning Indicators for Costly Asset Price Boom/Bust Cycles: A Role for Global Liquidity[J]. European Journal of Political Economy, 2011, 27(3): 520-533.
[23] Biau O, Angela D’Elia. Euro area GDP Forecasting Using Large Survey Datasets A Random Forest Approach[J]. Ecomod, 2010: 1-22.
[24] 郑挺国, 王霞. 中国经济周期的混频数据测度及实时分析[J]. 经济研究, 2013, 48(6): 58-70.
[25] 陈彦斌, 刘玲君, 陈小亮. 中国通货膨胀率预测——基于 LSTM 模型与 BVAR 模型的对比分析[J]. 财经问题研究, 2021(6): 18-29.
[26] Quinlan J R. C4.5: Programs for Machine Learning[M]. Morgan Kaufmann Publishers Inc. 1992.
[27] 周志华. 机器学习[M]. 第一版. 清华大学出版社, 2016: 85-88, 179-181.
[28] Breiman, Leo. Bagging Predictors[J]. Machine Learning, 1996, 24(2): 123-140.
[29] Breiman. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
[30] Xu L, Krzyzak A, Suen C Y. “Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition,” in IEEE Transactions on Systems, Man, and Cybernetics, 1992, 22(3): 418-435.
[31] 蒋锋, 张文雅. 机器学习方法在经济研究中的应用[J]. 统计与决策, 2022, 38(4): 43-49.

Construction of a Macroeconomic Leading Index System Based on Random Forest Algorithm

Zhao Yunfei[1] Lou Feng[2] Cheng Yuan[2]

(1. School of Applied Economics, University of Chinese Academy of Social Sciences;
2. Institute of Quantitative and Technological Economics, Chinese Academy of Social Sciences)

Abstract: Data from January 2001 to June 2022 are selected as samples to construct macroeconomic leading indices for Gross Domestic Product (GDP), price index, and exports, and we use data from July to December 2022 for out-of-sample prediction. The Random Forest (RF) index system constructed in this article is compared with existing indices. The results indicate the following: 1) Compared to traditional leading indicators, the RF leading index system better captures the trends of macroeconomic variables and fits historical economic turning points and volatility. 2) The RF index construction method is robust in terms of setting different leading times and selecting different forecasted variables within the sample. 3) The RF index successfully forecasts the trends and amplitudes of nominal GDP, Producer Price Index (PPI), and export indices in out-of-sample data. It also performs relatively well in forecasting the trends of real GDP and Consumer Price Index (CPI), although its amplitude predictions are less accurate. Therefore, the RF leading index system and method significantly improve the calculation methods of current indices, simplicity of operation, the robustness of time and variable selection, and forecast performance.

Key words: Random Forest; Economic Leading Index; Economic Forecast; Prosperity Index

（责任编辑：黄 煌）