研究与开发

基于电力大数据的中国工业增加值现时预测

彭放[1]，彭高群[1]，祁亚茹[1]，刘甜甜[1]，周晓磊[2]

（1. 国家电网有限公司大数据中心，北京 100031；2. 清华大学社会科学学院，北京 100084）

摘　要：工业增加值是衡量实体经济运行状况的重要指标，为充分挖掘电力数据在宏观经济现时预测中的价值，服务于政府政策制定和经济社会发展，以电力数据和传统统计数据为基础，采用 Bagging 和 Boosting 等算法对工业增加值进行了现时预测。结果表明：第一，传统统计数据可以显著提升 ARIMA 模型的预测效果；第二，电力数据的预测效果取决于电力指标的选择，选择恰当的电力指标有助于更及时、准确地预测工业增加值；第三，电力数据信息对当期工业增加值的预测能力可能会低于对未来期工业增加值的预测能力，这意味着电力数据更适用于超前的预测。

关键词：工业增加值；现时预测；电力数据

中图分类号：F224

文献标识码：A

doi: 10.11959/j.issn.1000−0801.2021143

Nowcasting of China’s industrial added value based on electric power big data

PENG Fang[1], PENG Gaoqun[1], QI Yaru[1], LIU Tiantian[1], ZHOU Xiaolei[2]

1. Big Data Center of State Grid Corporation of China, Beijing 100031, China
2. School of Social Sciences, Tsinghua University, Beijing 100084, China

Abstract: Industrial added value is an important indicator to measure the operation of the real economy. In order to fully mine the value of power data in the current macroeconomic nowcasting, so as to serve the government policy making, the Bagging and Boosting algorithms in machine learning were applied to nowcast industrial added value based on electric power data as well as traditional statistical data. Firstly, traditional statistical data can significantly improve the forecasting effect of the ARIMA model. Secondly, the nowcasting ability of electric power data depends on the selection of electric power indicators, and the proper electric power index is helpful to predict industrial added value more timely and accurately. Thirdly, the prediction ability of electric power data to the industrial added value in the current period may be lower than that in the future, which means the power data is more likely to be used to predict ahead of time.

Key words: industrial added value, nowcasting, electric power data

收稿日期：2020−10−12；修回日期：2021−06−20

通信作者：祁亚茹，xicaiqiyaru@163.com

基金项目：国家电网有限公司大数据中心科技项目（No.SGSJ0000FXJS2000098）

Foundation Item: The Technology Project Supported by Big Data Center of State Grid Corporation of China (No. SGSJ0000FXJS2000098)

1 引言

作为国民经济的“晴雨表”，电力数据可以及时、客观地反映经济发展的真实情况。伴随着电力信息化的发展，如何充分挖掘电力数据的价值服务于经济社会发展、政府政策制定成为关注的重点。工业增加值是衡量实体经济运行状况的重要指标，对其进行更加及时、准确地预测有助于政府更快、更准确地了解经济形势，为经济政策制定提供可靠的依据，同时也将有助于企业和个人准确把握经济形势并进行决策。在当前经济下行压力增加、不确定性加剧的环境下，及时准确地预测工业增加值显得尤为重要。为此，本文以工业增加的现时预测为出发点，充分挖掘电力大数据在现时预测中的作用。现时预测最初被用于气象领域，近年来被引入经济领域，是对当前经济形势的预测，其基本原理为在官方发布相关统计数据之前，充分挖掘先行高频数据的信息预测当前经济形势[1]。

当前关于中国工业增加值预测的研究相对较少，且主要基于所选择的若干传统统计指标进行，如利用社会消费品零食总额、广义货币数据对工业增加值进行预测[2]，基于构建的民营企业信用利差指数对工业增加进行预测[3]。在宏观经济传导机制愈发复杂、工业增加值影响因素更加多元的情形下，恰当的预测模型应当包含大量的解释变量以充分利用有价值的信息，仅利用少数几个变量较难获得精准的预测。此外，传统统计数据发布得滞后致使预测时难以利用经济活动的当前信息，而电力大数据与经济活动密切相关且实时可得，可提供经济活动的当前信息弥补传统统计数据的不足。因此，综合利用电力数据和传统统计数据可能得到更加准确的工业增加值预测结果。然而，这不可避免地引出工业增加值预测中传统数据和电力数据的关系问题，即工业增加值预测中电力数据是否可以替代传统数据，较之于成熟的统计数据电力数据的用处不大，两者是否可以相互补充。该问题同样也是大数据时代基于新兴数据进行宏观经济预测时可能碰到的问题，但当前却鲜有文献对此进行研究。本文设计了 5 类模型以探究工业增加值预测中统计数据与电力大数据的关系，同时采用多种预测方法进行预测以期获得更佳的预测效果。

2 研究设计

2.1 预测模型

为充分探究上述工业增加值现时预测中传统经济统计数据和电力数据的关系问题，设计两类数据结合的 5 种方式：其一，仅利用传统统计数据进行预测；其二，以传统经济统计数据预测为主，以电力数据预测为辅；其三，仅利用电力数据进行预测；其四，以电力数据预测为主，以统计数据预测为辅；其五，同时无差别地利用经济统计数据和电力数据进行预测。本文基于这 5 种结合方式构建了如下 5 类相应的预测模型。

模型 1：

$$\hat{y}_T = f(\{x_{1t},\cdots,x_{mt}\}_{t=1}^{T-1}) + \varepsilon_T \qquad (1)$$

本模型分析仅利用传统经济统计数据预测 y 的预测效果。模型中 $\{x_{1t},\cdots,x_{mt}\}_{t=1}^{T-1}$ 表示在 T 期预测时可用的信息集，由于经济统计数据的发布存在滞后性，所以信息集中仅包含 $T-1$ 期及其以前的统计数据信息。

模型 2：

$$\hat{y}_T = f(\{x_{1t},\cdots,x_{mt}\}_{t=1}^{T}) + f(\{z_{1t},\cdots,z_{nt}\}_{t=1}^{T} / \{x_{1t},\cdots,x_{mt}\}_{t=1}^{T-1}) + \varepsilon_T \qquad (2)$$

本模型利用传统经济统计数据和电力数据预测 y，在预测时以统计数据预测为主，以电力数据预测为辅。模型中 $\{z_{1t},\cdots,z_{nt}\}_{t=1}^{T}$ 表示在 T 期预测时可用的信息集，由于电力数据实时可得，所以预测第 T 期 y 值时可用到电力数据的当期值，即

第 T 期的值。$f(\{z_{1t},\cdots,z_{nt}\}_{t=1}^{T}/\{x_{1t},\cdots,x_{mt}\}_{t=1}^{T-1})$ 表示利用电力数据具有但不包含于统计数据中的信息进行预测。

模型 3：

$$\hat{y}_T = f(\{z_{1t},\cdots,z_{nt}\}_{t=1}^{T}) + \varepsilon_T \quad (3)$$

本模型分析仅利用电力数据预测 y 的预测效果。模型中 $\{z_{1t},\cdots,z_{nt}\}_{t=1}^{T}$ 表示在 T 期预测时可用的信息集，由于电力数据实时可得所以预测时可用到电力数据的当期值。

模型 4：

$$\hat{y}_T = f\{z_{1t},\cdots,z_{nt}\}_{t=1}^{T} + f(\{x_{1t},\cdots,x_{mt}\}_{t=1}^{T-1}/\{z_{1t},\cdots,z_{mt}\}_{t=1}^{T}) + \varepsilon_T \quad (4)$$

本模型利用传统经济统计数据和电力数据预测 y，在预测时以电力数据预测为主，以统计数据预测为辅。模型中 $\{z_{1t},\cdots,z_{nt}\}_{t=1}^{T}$ 表示在 T 期预测时可用的信息集，$f(\{x_{1t},\cdots,x_{nt}\}_{t=1}^{T}/\{z_{1t},\cdots,z_{mt}\}_{t=1}^{T-1})$ 表示在利用统计数据进行预测前先将统计数据中包含的电力数据信息去除。

模型 5：

$$\hat{y}_T = f(\{x_{1t},\cdots,x_{mt}\}_{t=1}^{T-1},\{z_{1t},\cdots,z_{mt}\}_{t=1}^{T}) + \varepsilon_T \quad (5)$$

模型 5 综合利用传统统计数据和电力数据预测工业增加值，其与模型 2、模型 4 的区别在于本模型在预测中无差别地对待两类数据。

2.2　预测方法

面对众多的解释变量，如何有效地利用其包含的信息是影响预测结果的关键问题。采用参考文献[4]提出的扩散指数（diffusion index，DI）算法、参考文献[5]提出的 Bagging 算法以及参考文献[6]提出的 Boosting 算法进行预测。

2.2.1　DI 算法

DI 算法为数据丰富环境下的常用预测方法，下面对扩散指数方法做简要介绍。对数据集 $\{x_{t1},\cdots,x_{tn},y_t\}_{t=1}^{T}$，首先，基于模型 $X_{tj} = f_t\boldsymbol{\Gamma}_j^{'} + \varepsilon_t$ 提取 $\{x_{t1},\cdots,x_{tn}\}_{t=1}^{T}$ 的共同因子 $f_t = (f_{1t},\cdots,f_{qt})$，然后利用提取的因子 f_t 构建如下模型并进行预测：

$$y_{t+h} = \beta_w^{'}w_t + \beta_f^{'}\hat{f}_t + \varepsilon_{t+h} \quad (6)$$

其中，y_t 为预测目标变量，$\hat{f}_t$ 为估计的因子向量，w_t 为可观测向量。对因子个数 q 的估计可利用如下信息准则[7]：

$$\hat{q}_N^T = \arg\min \mathrm{IC}_N^T(q) \quad (7)$$

其中：

$$\mathrm{IC}_N^T(q) = \log \mathrm{V}\left(q,\hat{f}^q,\hat{\Lambda}^q\right) + qp(N,T) \quad (8)$$

其中，$\mathrm{V}(q,\hat{f}^q,\hat{\Lambda}^q)$ 表示残差平方和的均值。$\hat{f}^q$和$\hat{\Lambda}^q$ 分别为因子个数 q 给定情形下的因子和因子载荷的估计值，$p(N,T)$ 为惩罚函数。

可以看出 DI 算法分为两步：首先，从众多变量构成的信息集中提取因子；然后，将因子纳入预测模型进行预测。第一步的因子提取既实现了降维又充分利用了大量变量的信息，但在提取因子时并未考虑信息集中的变量是否有助于预测，这可能导致纳入模型的因子不仅无助于预测反而降低了预测精准度。为此，采用 Bagging 算法[5]、Boosting 算法[6]解决上述问题以提高工业增加值的预测精度。

2.2.2　Bagging 算法

Bagging 为 bootstrap aggregation 的缩写，由 Breiman 提出[8]。该算法基于提取的 bootstrap 样本训练模型并进行预测，上述过程重复若干次后将所有预测值的均值作为最终预测结果。近年来，部分学者将其用于经济时间序列的预测[9-10]，本文参照参考文献[5]的做法进行预测，相应 Bagging 算法可表述为：

$$\hat{y}_{t+h}^{\mathrm{Bagging}} = w_t\hat{\beta}_w + \sum_{j=1}^{r}\psi(t_j)\hat{\beta}_{fj}f_{t,j} \quad (9)$$

其中，$\hat{\beta}_w$ 为 y_{t+h} 对 w_t 回归得到的估计值，$\hat{\beta}_{fj}$ 为 $y_{t+h} - w_t\hat{\beta}_w$ 对 $\hat{x}_{t-h,j}$ 回归得到的估计值，$t_j = \sqrt{T}\hat{\beta}_{fj}/S_{\mathrm{e}}$ 为 $\hat{\beta}_{fj}$ 的 t 统计量，S_{e} 为 Newey-West 标准误。ψ 为如下所指定的函数：

$$\psi(t) = 1 - \varnothing(t+c) + \varnothing(t-c) + t^{-1}[\phi(t-c) - \phi(t+c)] \quad (10)$$

其中，$\varnothing$为标准正态分布函数，ϕ为标准正态密度函数，c为预先设定的值。

2.2.3 Boosting 算法

Boosting 算法是一族可将弱学习器提升为强学习器的算法，一般可表示为如下形式：

$$\hat{y}^M = \sum_{m=1}^{M} \alpha_m f(X;\beta_m) \tag{11}$$

其中，α_m为权重，$f(X;\beta_m)$为定义在信息集X上函数。参考文献[11]提出的 Boosting 算法在每次迭代中仅利用一个变量，使得 Boosting 算法可适用于存在大量变量且数据满足独立同分布的场景，参考文献[6]进一步优化了参考文献[11]的算法，利用 Boosting 算法处理时间序列问题。采用参考文献[6]提出的 Component-Wise L2 Boosting 算法挑选用于预测的因子，若将 Boosting 算法作用于因子后得到的值记为$\hat{u}^m(\hat{f}_t)$，那么最终预测值为：

$$\hat{y}_{t+h}^{\text{Boosting}} = w_t\hat{\beta}_w + \hat{u}^m(\hat{f}_t) \tag{12}$$

3 实证分析

3.1 数据说明与处理

本文选择的数据分为两类：传统经济统计数据和电力数据。所有变量均采用同比增长率，时间跨度为 2014 年 1 月至 2017 年 12 月。传统统计数据主要包括居民消费价格指数、名义有效汇率、上证交易所和深证交易所各类指数、商业信心指数、贸易指数、货币供应、存款、宏观经济景气指数等 42 个变量，这些数据来源于 CEIC 数据库。电力数据为不同行业购电量数据的同比值，相关数据来源于国家电网大数据中心。

从信息集中提取因子要求信息集中的变量必须为平稳变量，为此本文采用常用的增广迪基−富勒（augmented Dickey-Fuller，ADF）单位根检验方法检验变量的平稳性。ADF 检验通过对如下 3 个模型的检验进行。

$$\Delta Y_t = \delta Y_{t-1} + \sum_{i=1}^{m} \beta_i \Delta Y_{t-i} + \varepsilon_t \tag{13}$$

$$\Delta Y_t = \alpha + \delta Y_{t-1} + \sum_{i=1}^{m} \beta_i \Delta Y_{t-i} + \varepsilon_t \tag{14}$$

$$\Delta Y_t = \alpha + \beta T + \delta Y_{t-1} + \sum_{i=1}^{m} \beta_i \Delta Y_{t-i} + \varepsilon_t \tag{15}$$

其中，Y_t表示被检验的时间序列，Δ为差分算子，即$\Delta Y_t = Y_t - Y_{t-1}$。上述 3 个模型的零假设均为$H0$：$\delta = 0$，即时间序列是不平稳的。上述 3 个模型只要有一个模型拒绝了零假设就认为时间序列是平稳的，如果 3 个模型都不能拒绝零假设则认为时间序列非平稳。对于被检验变量，若检验表明其不平稳，则对其取差分并再次检验其平稳性；若变量差分后仍不平稳，则再次进行差分。限于篇幅原因，本文不给出变量平稳性的检验结果及对非平稳变量的差分情况。

3.2 基准模型与效果评估

本文选择自回归滑动平均（auto-regressive integratedmoving average，ARIMA）模型作为基准模型，采用均方预测误差（mean squared forecast error，MSFE）测度模型的预测效果，MSFE 的计算式如下：

$$\text{MSFE} = \frac{1}{N}\sum_{t=1}^{N}\left(\hat{y}_t - y_t\right)^2 \tag{16}$$

其中，$\hat{y}_t$为被预测目标的预测值，y_t为预测目标的实际值，N为预测集所含样本数。进一步地本文用均方预测误差比值（ratio of mean squared forecast error，rMSFE）衡量预测模型的优劣。均方预测误差比值，即非基准模型均方预测误差与基准预测模型的比值，可表示如下：

$$\text{rMSFE} = \text{MSFE}_{\text{NBM}} / \text{MSFE}_{\text{BM}} \tag{17}$$

其中，MSFE_{NBM}为非基准模型的均方预测误差，MSFE_{BM}为基准预测模型的均方预测误差。rMSFE 小于 1，则说明非基准模型较基准模型有更好的预测效果，而且比值越小说明非基准模型的预测效果越好。

3.3　结果比较与分析

本文在估计模型和进行预测时采用了移动窗口和累积窗口两种方法。累积窗口法，即在下一期预测时，将前面所有样本（包括上一期）的真实值作为训练集放入估计模型中，训练集逐期增加。移动窗口法，就是固定训练集观察值的长度，每往后预测一期，训练集往后移动一期。本文选择 2014 年 1 月至 2017 年 6 月期间的数据作为训练集，选择 2017 年 7 月至 2017 年 12 月期间的数据作为预测集。

为充分探究电力数据在工业增加值预测中的作用，报告基于电力数据指标不同分类与传统统计数据组合的预测效果进行了分析，表 1 至表 9 分别呈现了不同模型与数据组合预测工业增加值的 rMSFE。考虑电力数据实时可得的特点，本文采用两种方式利用电力数据信息：其一，仅利用电力数据当期信息，即表 1 至表 9 中数据无滞后；其二，不仅利用电力数据的当期信息还利用其滞后信息，即表 1 至表 9 中数据有滞后。

表 1 呈现了基于居民和企业总用电数据与统计数据现时预测工业增加值的 rMSFE，表 2 呈现了利用全行业用电一级指标数据与统计数据现时预测工业增加值的 rMSFE。由表 1 和表 2 可以发现以下结论。

（1）对模型 1，TMSFE 均小于 1，这意味着利用 ARIMA 模型的预测结果并不是最优的，基于大量传统统计数据的预测效果要优于 ARIMA 模型。

（2）对模型 2 到模型 5，与仅利用电力数据当期信息的预测效果相比，利用电力数据当期和滞后信息的预测效果更好。为此本文仅就存在滞后信息的情况进行分析。

表 1　基于居民和企业总用电数据与统计数据预测工业增加的 rMSFE

方法		移动窗口					累计窗口				
		模型 1	模型 2	模型 3	模型 4	模型 5	模型 1	模型 2	模型 3	模型 4	模型 5
有滞后	DI	0.596 0	2.715 7	1.632 2	1.898 0	0.594 4	0.590 3	2.654 6	1.525 1	1.947 9	0.579 1
	Bagging	0.605 4	2.569 4	1.677 8	2.077 2	0.621 1	0.593 5	2.546 2	1.541 1	2.016 5	0.577 5
	Boosting	0.912 0	3.162 2	1.202 9	1.319 3	0.899 5	0.928 5	2.223 9	1.089 6	1.389 5	0.918 9
无滞后	DI	0.596 0	1.563 0	1.448 2	1.600 8	0.599 1	0.590 3	1.537 9	1.435 8	1.596 9	0.595 1
	Bagging	0.605 4	1.537 9	1.462 4	1.627 5	0.611 7	0.593 5	1.509 2	1.435 8	1.585 8	0.595 1
	Boosting	0.912 0	1.624 4	1.369 6	1.278 4	0.908 9	0.928 5	1.761 3	1.402 3	1.442 2	0.926 9

表 2　基于全行业用电一级指标数据与统计数据预测工业增加的 rMSFE

方法		移动窗口					累计窗口				
		模型 1	模型 2	模型 3	模型 4	模型 5	模型 1	模型 2	模型 3	模型 4	模型 5
有滞后	DI	0.596 0	1.158 9	0.748 5	1.390 1	1.118 0	0.590 3	2.030 9	0.574 3	1.472 5	0.947 6
	Bagging	0.605 4	1.135 3	0.773 7	1.369 6	1.116 5	0.593 5	1.984 6	0.582 3	1.437 4	0.950 8
	Boosting	0.912 0	1.366 5	0.943 5	1.122 7	1.280 0	0.928 5	1.900 0	0.847 1	1.439 0	1.233 2
无滞后	DI	0.596 0	1.622 8	0.891 6	1.577 2	0.596 0	0.590 3	1.268 3	0.828 0	0.893 4	0.591 9
	Bagging	0.605 4	1.517 4	0.860 1	1.382 2	0.613 3	0.593 5	1.233 2	0.749 8	0.740 2	0.593 5
	Boosting	0.912 0	1.226 5	0.838 1	1.003 2	0.910 5	0.928 5	1.515 6	0.855 1	0.457 9	0.930 1

（3）对模型 3，其在表 2 中的预测效果要优于在表 1 中的预测效果，而且前者的效果要优于 ARIMA 模型，后者的预测效果与 ARIMA 模型相比较差。这意味着相对于居民和企业总用电数据，电力全行业一级指标用电数据包含更多有助于工业增加值预测的信息。

（4）对模型 2，可以发现利用电力全行业一级指标以及居民和企业总用电量的预测效果均不优于模型 1，而且利用电力数据全行业一级指标的预测效果与利用居民和企业总用电量的预测效果相比更优，这意味着居民和企业总用电量包含更多无助于工业增加值预测的信息。

（5）对模型 4，可以发现其大多数情况下预测效果比模型 3 差，这表明在剔除相应电力数据信息后，统计数据剩余信息含有较多无助于工业增加值预测的噪音。对模型 5，由于无差别地利用两类数据，且表 1、表 2 所用电力数据指标个数分别为 1 和 8，基于因子提取的机理可推断模型 5 的预测效果在表 1 中其接近于模型 1，在表 2 中其接近于模型 3。

上述分析表明，就工业增加值的预测而言，电力数据全行业一级指标比居民和企业用电总量指标含有更多有用的信息，但仍包含较多无助于预测的信息，为此接下来的分析将减少电力数据指标涉及的行业。表 3、表 4 分别给出了利用第二产业总用电数据与传统统计数据、工业总用电数据与传统统计数据现时预测工作增加值的 rMSFE。观察表 3、表 4 可以发现以下结论。

（1）对模型 2 到模型 5，较之于仅利用电力数据当期信息的预测效果，利用电力数据当期和滞后信息的预测效果更好。这意味着第二产业总用电数据、工业总用电数据的历史信息较之于当期信息具有更好的预测作用，具有领先性。故下面

表 3　基于第二产业总用电数据与统计数据工业增加的 rMSFE

方法		移动窗口					累计窗口				
		模型 1	模型 2	模型 3	模型 4	模型 5	模型 1	模型 2	模型 3	模型 4	模型 5
有滞后	DI	0.596 0	2.133 8	0.418 3	1.520 6	0.589 7	0.590 3	1.577 8	0.408 4	1.660 7	0.579 1
	Bagging	0.605 4	2.102 4	0.405 7	1.165 2	0.605 4	0.593 5	1.550 7	0.408 4	1.322 5	0.574 3
	Boosting	0.912 0	1.108 6	0.710 8	0.897 9	0.894 7	0.928 5	1.025 8	0.743 4	1.054 5	0.950 8
无滞后	DI	0.596 0	2.022 2	1.426 2	1.718 7	0.600 7	0.590 3	1.994 2	1.344 9	1.877 7	0.593 5
	Bagging	0.605 4	1.967 2	1.442 0	1.718 7	0.611 7	0.593 5	1.938 3	1.344 9	1.802 7	0.595 1
	Boosting	0.912 0	1.953 0	1.341 3	1.191 9	0.907 3	0.928 5	2.136 2	1.348 1	1.435 8	0.925 3

表 4　基于工业总用电数据与统计数据工业增加的 rMSFE

方法		移动窗口					累计窗口				
		模型 1	模型 2	模型 3	模型 4	模型 5	模型 1	模型 2	模型 3	模型 4	模型 5
有滞后	DI	0.596 0	2.146 4	0.407 3	1.567 8	0.589 7	0.590 3	1.506 0	0.397 2	1.684 7	0.579 1
	Bagging	0.605 4	2.122 8	0.386 8	1.169 9	0.605 4	0.593 5	1.494 8	0.392 5	1.236 4	0.574 3
	Boosting	0.912 0	1.250 1	0.684 0	0.968 6	0.894 7	0.928 5	1.037 0	0.730 7	0.901 4	0.950 8
无滞后	DI	0.596 0	1.976 6	1.430 9	1.681 0	0.600 7	0.590 3	1.970 2	1.354 4	1.855 4	0.593 5
	Bagging	0.605 4	1.921 6	1.446 7	1.690 4	0.611 7	0.593 5	1.914 4	1.354 4	1.783 6	0.595 1
	Boosting	0.912 0	1.913 7	1.341 3	1.173 1	0.907 3	0.928 5	2.113 8	1.351 3	1.421 4	0.925 3

仅就存在滞后的情况进行分析。

（2）对模型 3，在表 3、表 4 中 rMSFE 均小于 1，这表明其预测效果均优于 ARIMA 模型。此外，由模型 3 在表 4 中具有更好的预测效果可知，工业总用电数据比第二产业总用电数据包含更多有助于工业增加值现时预测的信息。

（3）比较模型 1 和模型 3 可知，模型 3 具有更好的预测效果，这表明较之于统计数据，第二产业总用电数据或者工业总用电数据含有更多有助于工业增加值预测的信息。

基于上述分析，工业总用电数据相对于第二产业总用电数据含有更多有助于工业增加值现时预测的信息。考虑工业部门包含不同行业，接下来探究工业部门中不同行业用电数据现时预测工业增加值的能力，预测结果如表 5 至表 9 所示。表 5、表 6 分别给出了基于工业部门用电二级指标数据与统计数据、工业部门用电一级指标数据与统计数据的预测效果。观察表 5、表 6 可以发现以下结论。

（1）对模型 2 到模型 5，用电力当期及滞后期数据的预测效果更好。这同样表明相关电力数据的历史信息比当期信息具有更好的预测作用，所以本文仅就存在滞后的情况进行分析。

（2）比较模型 3 与模型 1，显然前者的预测效果优于后者。这意味着较之于统计数据，表 5、表 6 所用电力数据含有更多有助于工业增加值预测的信息。这点通过比较模型 2 与模型 1、模型 4 与模型 3 的预测效果也可以看出。

对比表 4、表 5、表 6 中模型 3 的预测效果可以发现，尽管都利用了工业部门的电力数据信息，但表 6 中模型 3 的预测效果却更好，即利用工业部门用电一级指标数据的预测效果要优于利用工

表 5　基于工业部门用电二级指标数据与统计数据工业增加的 rMSFE

方法		移动窗口					累计窗口				
		模型 1	模型 2	模型 3	模型 4	模型 5	模型 1	模型 2	模型 3	模型 4	模型 5
有滞后	DI	0.596 0	0.550 4	0.404 1	1.328 7	0.520 5	0.590 3	0.449 9	0.473 8	1.054 5	0.504 1
	Bagging	0.605 4	0.566 1	0.459 2	1.383 8	0.536 2	0.593 5	0.480 2	0.488 2	1.021 0	0.499 3
	Boosting	0.912 0	0.789 4	0.652 6	1.163 6	0.795 7	0.928 5	0.721 1	0.727 5	0.907 7	0.877 4
无滞后	DI	0.596 0	1.765 9	1.638 5	2.179 4	0.555 1	0.590 3	2.054 8	1.555 5	2.239 9	0.529 7
	Bagging	0.605 4	1.740 7	1.654 2	2.155 9	0.536 2	0.593 5	2.002 2	1.555 5	2.190 4	0.532 8
	Boosting	0.912 0	1.747 0	1.298 9	1.182 5	0.813 0	0.928 5	1.730 9	1.328 9	1.233 2	0.818 4

表 6　基于工业部门用电一级指标数据与统计数据工业增加的 rMSFE

方法		移动窗口					累计窗口				
		模型 1	模型 2	模型 3	模型 4	模型 5	模型 1	模型 2	模型 3	模型 4	模型 5
有滞后	DI	0.596 0	0.355 4	0.336 5	0.945 1	0.561 4	0.590 3	0.362 1	0.366 9	0.705 1	0.555 2
	Bagging	0.605 4	0.366 4	0.350 7	0.893 2	0.575 5	0.593 5	0.366 9	0.365 3	0.673 2	0.555 2
	Boosting	0.912 0	0.798 8	0.655 7	0.831 8	0.880 6	0.928 5	0.810 4	0.634 9	0.842 3	0.887 0
无滞后	DI	0.596 0	1.781 6	1.143 2	1.446 7	0.589 7	0.590 3	1.624 1	0.914 1	1.504 4	0.580 7
	Bagging	0.605 4	1.769 0	1.152 6	1.405 8	0.594 4	0.593 5	1.571 4	0.914 1	1.439 0	0.582 3
	Boosting	0.912 0	1.451 4	0.622 7	0.652 6	0.899 5	0.928 5	1.600 1	0.843 9	0.917 3	0.918 9

表 7　基于采矿业总用电数据与统计数据工业增加的 rMSFE

方法		移动窗口					累计窗口				
		模型 1	模型 2	模型 3	模型 4	模型 5	模型 1	模型 2	模型 3	模型 4	模型 5
有滞后	DI	0.596 0	0.245 3	0.086 5	0.240 6	0.580 2	0.590 3	0.293 5	0.118 1	0.191 4	0.571 1
	Bagging	0.605 4	0.256 3	0.092 8	0.209 1	0.592 8	0.593 5	0.299 9	0.129 2	0.183 5	0.571 1
	Boosting	0.912 0	0.283 0	0.209 1	0.283 0	0.899 5	0.928 5	0.374 9	0.454 7	0.424 4	0.910 9
无滞后	DI	0.596 0	0.506 3	0.449 7	0.493 8	0.589 7	0.590 3	0.433 9	0.406 8	0.422 8	0.583 9
	Bagging	0.605 4	0.504 8	0.452 9	0.489 0	0.599 1	0.593 5	0.433 9	0.406 8	0.429 1	0.585 5
	Boosting	0.912 0	0.613 3	0.597 5	0.608 5	0.905 7	0.928 5	0.646 1	0.599 8	0.654 1	0.923 7

表 8　基于制造业总用电数据与统计数据工业增加的 rMSFE

方法		移动窗口					累计窗口				
		模型 1	模型 2	模型 3	模型 4	模型 5	模型 1	模型 2	模型 3	模型 4	模型 5
有滞后	DI	0.596 0	0.638 4	0.402 6	1.926 3	0.583 4	0.590 3	1.175 8	0.433 9	1.952 7	0.574 3
	Bagging	0.605 4	0.594 4	0.391 5	1.567 8	0.600 7	0.593 5	1.137 5	0.459 5	1.609 7	0.574 3
	Boosting	0.912 0	0.591 2	0.641 6	1.171 5	0.894 7	0.928 5	0.619 0	0.749 8	1.040 2	0.904 6
无滞后	DI	0.596 0	1.758 0	1.258 0	1.608 6	0.599 1	0.590 3	1.748 5	1.199 7	1.662 3	0.591 9
	Bagging	0.605 4	1.712 4	1.272 1	1.600 8	0.605 4	0.593 5	1.692 7	1.199 7	1.612 9	0.595 1
	Boosting	0.912 0	1.602 3	1.251 7	1.047 3	0.907 3	0.928 5	1.783 6	1.265 1	1.183 7	0.925 3

表 9　基于电力、燃气及水的生产和供应业总用电数据与统计数据工业增加的 rMSFE

方法		移动窗口					累计窗口				
		模型 1	模型 2	模型 3	模型 4	模型 5	模型 1	模型 2	模型 3	模型 4	模型 5
有滞后	DI	0.596 0	0.852 3	0.844 4	1.294 1	0.588 1	0.590 3	0.855 1	0.912 5	1.102 4	0.585 5
	Bagging	0.605 4	0.833 4	0.819 3	1.231 2	0.600 7	0.593 5	0.840 7	0.922 1	1.107 2	0.587 1
	Boosting	0.912 0	1.506 4	1.045 7	1.339 7	0.907 3	0.928 5	1.558 6	1.121 5	1.368 8	0.918 9
无滞后	DI	0.596 0	0.511 1	0.669 9	0.506 3	0.594 4	0.590 3	0.465 8	0.595 1	0.480 2	0.587 1
	Bagging	0.605 4	0.504 8	0.674 6	0.567 7	0.602 3	0.593 5	0.469 0	0.595 1	0.488 2	0.588 7
	Boosting	0.912 0	0.888 4	0.918 3	0.835 0	0.912 0	0.928 5	0.885 4	0.891 8	0.816 8	0.928 5

业部门总用电数据及工业部门用电二级指标数据的预测效果。结合因子分析的机理，认为这可能源于工业部门中部分行业的用电数据无助于工业增加预测。故接下来将统计数据分别与表 6 中所用工业部门用电一级指标（即矿业、制造业、电力、燃气及水的生产和供应业 3 个行业各自用电总量数据）结合进行工业增加值现时预测，相应的预测结果见表 7、表 8、表 9。通过比较可以发现以下结论。

（1）对模型 2 到模型 5 的预测效果，在表 7 和表 8 中利用电力数据当期及滞后期信息的预测效果较之于仅利用电力数据当期信息的预测效果更优，但在表 9 中则仅利用电力数据当期信息时的预测效果更好。故接下来的分析对表 7

和表 8 仅讨论利用电力数据当期及滞后信息的情形，对表 9 仅讨论利用电力数据当期信息时的情形。

（2）对模型 3，其在表 7、表 8、表 9 中的预测效果依次递减；模型 3 在表 6 中的预测效果明显地弱于其在表 7 中的预测效果，但优于其在表 8、表 9 中的预测效果。

（3）对于模型 1 与模型 3，可以发现在表 9 中前者的预测效果明显优于后者，而在表 6、表 7 及表 8 中后者的预测效果优于前者，特别是在表 7 中更是如此。

上述关于模型 3、模型 1 在表 6、表 7、表 8 及表 9 中预测效果的分析表明利用电力、燃气及水的生产和供应业电力数据与统计数据的组合可能不是现时预测工业增加值的最佳组合。此外，在表 7 中模型 2 的预测效果优于模型 1，而在表 8 中除 Boosting 算法外模型 2 的预测效果并不优于模型 1，这在一定程度上表明采矿业总体用电数据去除统计数据信息后仍含有较多有助于工业增加值预测的信息，而制造业总体用电数据在去除统计数据信息后所含有助于工业增加值预测的信息相对较少。

以上部分基于表 1 至表 9 中的 rMSFE 分析了 DI、Bagging、Boosting 算法与两类数据不同结合方式的预测效果，图 1 给出了上述各表中移动窗口下 3 种算法各自的最佳预测值曲线，以更加直观地呈现不同预测方法及数据结合方式的预测效果。观察图 1 可以发现，表 3、表 4、表 6、表 7 中各预测方法的预测值曲线与工业增加值实际值曲线趋势一致，预测残差相对较小，而且就预测值曲线趋势和预测残差来看，基于 DI 算法的预测曲线相对更优。

综合上述分析可以发现，就工业增加值的现时预测而言，第二产业、工业、采矿业、制造业的总体用电数据均有较好的预测效果，而且采矿业总体用电数据与统计数据的组合是相对更优的数据组合。就预测方法而言，DI 的预测效果要优于 Bagging、Boosting，这或许与所用数据有关。此外，较之于仅仅利用电力数据当期信息，利用电力数据当期及滞后期信息可实现更好的工业增加值现时预测效果。

4　结束语

伴随着人类社会迈入大数据时代，部分学者尝试利用新兴大数据进行宏观经济预测。新兴大数据与传统经济统计数据各具优劣，但现有的宏观经济预测研究或者侧重对一类数据的挖掘应用，或者对两类数据无区别地使用，涉及对这两类数据关系的探讨及区别对待的研究较少。电力数据作为经济发展的“晴雨表”，与经济活动密切相关，但鲜有文献探究电力数据在宏观经济预测中的价值。本文提出了电力大数据与经济统计数据在宏经济预测中相结合的 5 种方式，采用 DI、Bagging、Boosting 等大数据预测算法，对工业增加值这一重要宏观经济指标进行了预测，探究了在工业增加值预测中如何有效地利用电力大数据与经济统计数据两类信息以获得更加及时、准确的预测结果。研究结果表明，基于选择出的恰当电力指标，可以得到更加及时、准确的工业增加值预测结果。本文的研究可为宏观经济预测中如何有效利用新兴数据和经济统计数据两类信息提供有益的借鉴与启示。

参考文献：

[1] 尹德才, 张文. 当前宏观经济现时预测研究综述[J]. 统计与信息论坛, 2020, 35(1): 121-128.
YIN D C, ZHANG W. A review of current macroeconomic nowcasting research[J].Statistics & Information Forum, 2020, 35(1): 121-128.

[2] 顾光同, 许冰. 中国工业增加值的半月预报: 基于宏观月度数据[J]. 系统工程理论与实践, 2018, 38(8): 1983-1993.
GU G T, XU B. Half-month forecast of China’s industrial added value: based on the macro monthly data[J]. Systems Engineering-Theory & Practice, 2018, 38(8): 1983-1993.

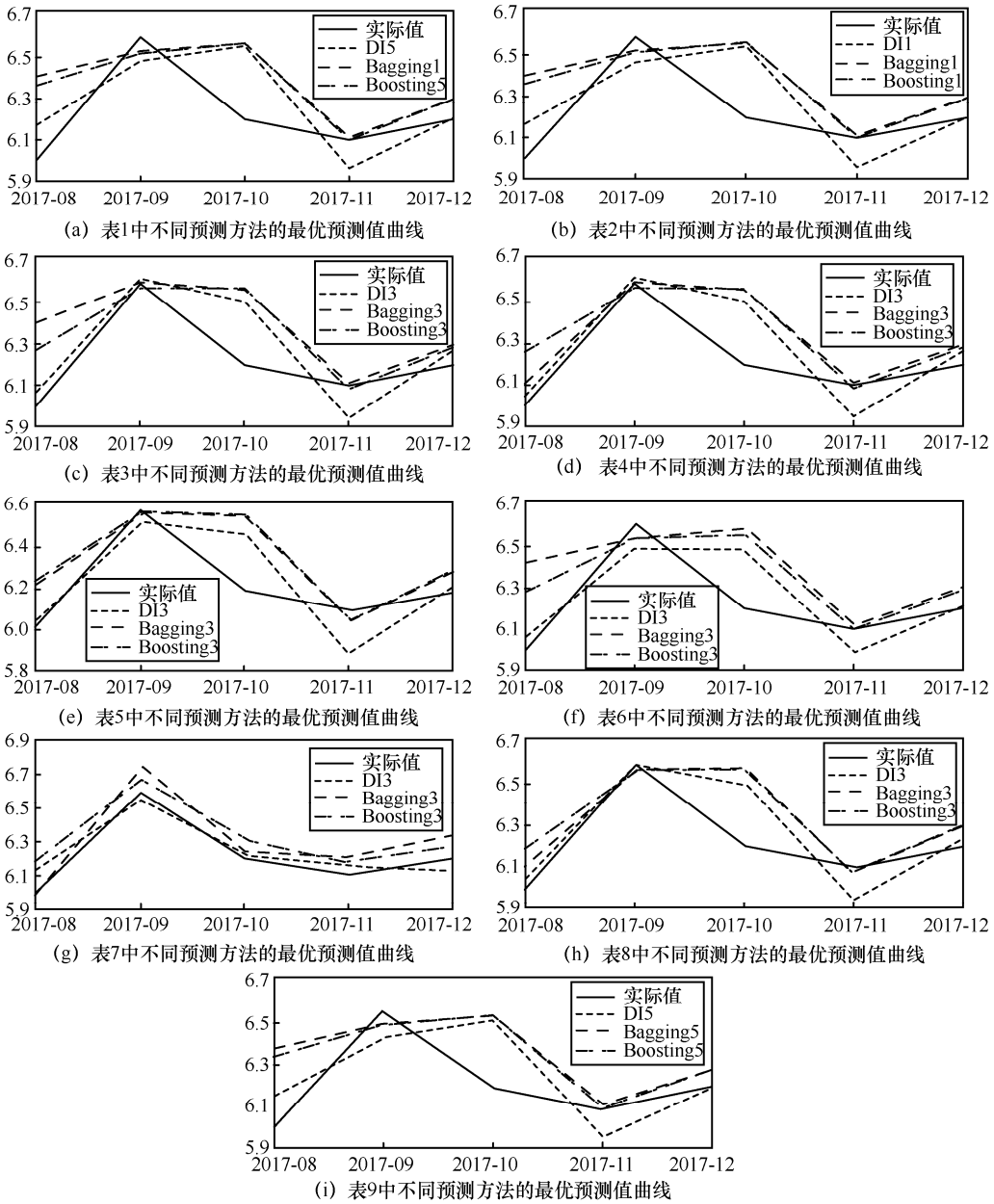

图 1　不同层级电力数据下 DI、Bagging 及 Boosting 方法的最优预测值

[3] 王雷, 聂常虹. 中国债券利差对宏观经济指标的预测能力研究[J]. 管理评论, 2019, 31(5): 66-76.
WANG L, NIE C H. Research on how predictive Chinese bond spreads is of macroeconomic indices[J]. Management Review, 2019, 31(5): 66-76.

[4] STOCK J H, WATSON M W. Forecasting using principal components from a large number of predictors[J]. Journal of the American Statistical Association, 2002, 97(460): 1167-1179.

[5] STOCK H, WATSON, W. Generalized shrinkage methods for forecasting using many predictors[J]. Journal of Business and Economic Statistics, 2012, 30(4): 481-493.

[6] BAI J, NG S. Boosting diffusion indices[J]. Journal of Applied Econometrics, 2009, 24(4): 607-629.

[7] BAI J, NG S. Determining the number of factors in approximate factor models[J]. Econometrica, 2002, 70(1): 191-221.

[8] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[9] INOUE A, KILIAN L. How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation[J]. Journal of the American Statistical Association, 2008,

103(482): 511-522.

[10] KIM H H, SWANSON N R. Forecasting financial and macroeconomic variables using data reduction methods: new empirical evidence[J]. Journal of Econometrics, 2014(178): 352-367.

[11] BÜHLMANN P, YU B. Boosting with the L_2 loss: regression and classification[J]. Journal of the American Statistical Association, 2003(98): 324-339.

[作者简介]

彭放（1987− ），男，国家电网有限公司大数据中心高级工程师、数据分析中心经营处处长，主要研究方向为经营管理和智慧政务大数据。

彭高群（1984− ），女，国家电网有限公司大数据中心数据分析中心经营处副处长，主要研究方向为智慧政务大数据。

祁亚茹（1990− ），女，国家电网有限公司大数据中心中级经济师，主要研究方向为智慧政务大数据。

刘甜甜（1992− ），女，国家电网有限公司大数据中心注册会计师，主要研究方向为财务、审计方向大数据分析。

周晓磊（1991− ），女，清华大学博士生，主要研究方向为宏观经济、经济大数据分析。