基于大数据的碳价预测

王 娜

内容提要:为了研究大数据是否有助于预测碳排放权价格,本文讨论了结构化数据和非结构化信息对预测碳价所起的作用。结构化数据选取国际碳现货价格、碳期货价格和汇率,非结构化信息选择百度搜索指数和媒体指数。考虑到当解释变量很多时,平等对待每一个解释变量是不合理的,本文提出了网络结构自回归分布滞后(ADL)模型,在参数估计和变量选择的同时兼顾了解释变量之间的网络关系。实证分析表明,网络结构 ADL 模型明显优于其他模型,可以获得较高的预测准确性,更适合基于大数据的预测。

关键词:大数据;网络结构;碳价预测

DOI:10.19343/j.cnki.11-1302/c.2016.11.008

中图分类号:C812　　**文献标识码**:A　　**文章编号**:1002-4565(2016)11-0056-07

Forecasting of Carbon Price Based on Big Data

Wang Na

Abstract:This paper analyzes the effect of structured data and unstructured information on carbon price forecasting in order to learn if big data can help us predict carbon price. We choose international carbon spot price, carbon futures price and exchange rate as structured data, Baidu search index and media index as unstructured information. Considering that it is not reasonable when there are a lot of explanatory variables by taking every variables equally, we proposes network autoregressive distributed lag(ADL) model, taking variables as a network when we estimate parameters and choose variables. The empirical results show that network ADL model is better than other models, which can get high accuracy, and it is more suitable for predicting based on big data.

Key words:Big Data; Network; Carbon Price Forecasting

中国是能源生产消费大国,为了共同控制温室气体的排放(简称"碳排放"),中国签署了《京都议定书》和《巴黎协定》,"十三五"规划建议提出的"五大发展"理念也将绿色发展放在了第三位。2013 年 6 月 18 日,深圳启动了首个碳排放权交易试点,随后,上海、北京、广东、天津、湖北、重庆等省市也先后启动了碳排放权交易试点,截止 2015 年 7 月底,全国共启动了 7 个碳排放权交易试点,累计成交额达 19.5 亿元。《中美元首气候变化联合声明》提出,中国将于 2017 年启动全国统一碳排放交易平台,用更完善的交易体制帮助实现绿色发展。

随着国内外碳交易市场的日趋成熟,对碳价的关注程度也越来越高,但对碳价预测的文献却很少,由于中国的碳交易市场起步较晚且尚未成熟,对中国碳价的预测就更屈指可数了。

一、文献综述

Zhu 和 Wei(2013)[1]把自回归融合滑动平均模型(ARIMA)和最小二乘支持向量机(LSSVM)结合,提出了兼顾线性和非线性的复合预测方法,并对欧盟气候交易所(ECX)的 DEC10 碳期货合约和 DEC12 碳期货合约进行了预测,取得了不错的预测效果。Li 和 Lu(2015)[2]利用深圳、上海、北京、广东和天津 5 个试点的价格数据,把经验模态分解(EMD)算法和广义自回归条件异方差(GARCH)模型结合,预测了 5 个试点 2016 年以后的碳价,为建设统一碳交易市场提出了一个价格参考区间,文章给出每吨 CO_2 当量的价格参考区间为 30 ~ 50 元。然而,这些文献都忽略了大数据因素对碳价的影响。

Armah(2013)[3]指出,大数据是指巨大多样的

数据集,包括结构化数据和非结构化数据。在网络技术发达的今天,有很多利用大数据进行的预测。Choi 和 Varian(2009)[4]发现失业和福利相关的搜索可以提高预测领取失业保险的准确性。Askitas 和 Zimmermann(2009)[5]发现互联网搜索能够预测劳动力市场。Choi 和 Varian(2012)[6]用"卡车及 SUV"和"汽车保险"关键词搜索指数预测汽车销售,用"工作"和"福利及失业"关键词搜索指数预测首次申请失业救济的人数,用"度假胜地"关键词搜索指数预测旅游目的地计划,用"犯罪与司法"、"卡车与 SUV"和"混合及替代燃料汽车"关键词搜索指数预测消费者信心。Schlegel(2014)[7]讨论了如何利用大数据进行预测分析,以管理供应链风险。刘涛雄和徐晓飞(2015)[8]使用了 PC 端百度搜索指数讨论了互联网搜索行为对宏观经济预测产生的影响,得出非结构化数据有助于提高预测宏观经济的准确性,但不能替代政府统计数据,且要使用合适的预测模型,由此提出了"两步法",即先使用政府统计数据进行初步预测,再加入百度搜索指数。虽然取得了不错的预测效果,但由于此文的目的是为了检测互联网搜索行为是否能够帮助预测宏观经济,所以作者限定只能新增 1 个搜索行为变量,没有对增加更多的非结构化信息进行详细论证,不能直接应用。而且,非结构化数据只使用了 PC 端百度搜索指数,具有局限性。更重要的是,文章没有就高维变量选择方法对预测产生的影响进行讨论,忽略了此因素的重要性。

本文将利用结构化的我国官方统计数据及非结构化的在线大数据对碳价进行预测,观察大数据对碳价预测所起的作用。本文选取的在线大数据不仅包括了代表 PC 端和移动端网民对碳价关心程度的百度整体搜索指数,还包括了反映网络媒体对碳价关注程度的媒体指数,力求用更全面的数据信息预测碳价。在借鉴"两步法"思想的同时,不限制新增变量的个数,旨在选出最优模型。

事实上,在使用大数据进行预测时,高维变量选择方法起着至关重要的作用,本文将对此进行深入研究。注意到,基于大数据的预测问题的解释变量都非常多,此时部分变量之间会存在相互影响的关系。Huang 等(2011)[9]指出把解释变量的相互影响关系处理成复杂网络将会提高变量选择和预测效果。方匡南等(2016)[10]考虑到变量间的网络结构关系,提出了网络结构 Logistic 模型,并将其应用到企业信用风险预警中,取得了良好的效果。本文结合时间序列的特殊性,提出了适合于大数据的碳价预测的网络结构 ADL 模型,在参数估计和模型选择的同时兼顾变量之间的网络关系,显著地提高了预测的精准度,为应用大数据预测碳价提供了高效可行的方法。

二、模型介绍

本文研究的被解释变量为我国碳交易试点最新的碳排放权价格,记为 y_t。解释变量包括 y_t 的 P 阶滞后,结构化的官方统计指标 $z^1_{j,t}, j = 1,\cdots,N_1$ 的 P 阶滞后,以及非结构化信息 $z^2_{j,t}, j = 1,\cdots,N_2$ 的 P 阶滞后,非结构化信息包含百度整体搜索指数和媒体指数。由于解释变量较多,为了防止解释变量过多带来"维数灾难",需要进行变量选择,本文将采取 ADL 模型与高维变量选择模型结合的方式进行碳价预测。具体模型如下:

(一)Lasso-ADL 模型 1

令:

$x_t = (y_{t-1},\cdots,y_{t-P},z^1_{1,t-1},\cdots,z^1_{1,t-P},\cdots,z^1_{N_1,t-1},\cdots,z^1_{N_1,t-P})$。

则 ADL 模型可写成:

$$y_t = \alpha_0 + \beta' x_t + \varepsilon_t = \alpha_0 + \sum_{p=1}^{P}\alpha_p y_{t-p} + \sum_{j=1}^{N_1}\sum_{p=1}^{P}\varphi^1_{jp} z^1_{j,t-p} + \varepsilon_t \quad (1)$$

本模型只将被解释变量的自身信息和结构化数据放入模型,主要是为了考察没有加入非结构化数据时的预测能力,此模型简称为"ADL 模型 1"。

考虑到 Lasso 的稳定性和准确性[11],将使用 Lasso 对 ADL 模型 1 进行变量选择。Lasso 定义如下:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\{\| y - \sum_{j=1}^{N} x_j\beta_j \|^2 + \lambda\sum_{j=1}^{N}|\beta_j|\} \quad (2)$$

其中,$N = (1 + N_1) \times P$,目标函数的第 1 项为残差平方和,使用的是最小二乘法。第 2 项为 Lasso 惩罚项,其中 λ 为非负调和参数,λ 越小,惩罚越小,被剔除掉的变量就越少,λ 越大,被剔除的变量则越多。通过控制 λ 的大小决定被保留变量的个数,使用交叉验证法选择 λ。本文使用 LARS[12]算法实现 Lasso。

(二) Lasso-ADL 模型 2

令:

$$Y_t = y_t - \hat{\alpha}_0 - \hat{\alpha} y - \hat{\varphi}^1 z^1$$

$$x_t = (z^2_{1,t-1}, \cdots, z^2_{1,t-P}, \cdots, z^2_{N_2,t-1}, \cdots z^2_{N_2,t-P})$$

其中, $y = (y_{t-1}, \cdots, y_{t-P})$, 代表被解释变量的自身信息。$\hat{\varphi}^1$、$\hat{\alpha}_0$、$\hat{\alpha}$ 为 Lasso-ADL 模型 1 中参数 $\varphi^1 = \{\varphi^1_{11}, \cdots, \varphi^1_{N_1 \times P}\}$、$\alpha_0$、$\alpha = (\alpha_1, \cdots, \alpha_P)$ 的估计值。$z^1 = (z^1_{1,t-1}, \cdots, z^1_{1,t-P}, \cdots, z^1_{N_1,t-1}, \cdots, z^1_{N_1,t-P})$, 代表结构化信息,则此时的 ADL 模型可写成:

$$Y_t = \alpha_0 + \beta' x_t + \varepsilon_t = \sum_{l=1}^{N_2} \sum_{p=1}^{P} \varphi^2_{lp} z^2_{l,t-p} + \varepsilon_t \quad (3)$$

本模型在 Lasso-ADL 模型 1 的基础上,加入非结构化信息,主要考察非结构化信息是否能够帮助预测碳价,此模型简称为"ADL 模型 2",同样使用 Lasso 对此模型进行变量选择,其中, $N = N_2 \times P$。本部分借鉴的是刘涛雄和徐晓飞(2015)[8]"两步法"的思想,不同的是,原文献限制只能新增 1 个搜索行为变量,本文对此不进行限制。

(三) 网络结构 ADL 模型

在 ADL 模型 2 中,解释变量非常多,部分解释变量之间可能会相互影响,需要考虑解释变量之间的网络关系,把这种兼顾变量网络结构的 ADL 模型称为网络结构 ADL 模型。定义如下:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \{ \| Y - \sum_{j=1}^{N} x_j \beta_j \|^2 + \lambda \sum_{j=1}^{N} \omega_j |\beta_j| \} \quad (4)$$

其中, $N = N_2 \times P$, 这是对 Lasso-ADL 模型 2 的改进,与 Lasso 不同的是,这里多了一个 ω。$\omega = (\omega_1, \cdots, \omega_N)^T$ 为惩罚项的权重向量,权重值由解释变量的网络结构决定。网络中影响强度越大的变量越重要,这种变量应该施以更小的惩罚,以便更好地保留住这些重要变量,对于影响强度不高的变量,应对其施以较大的权重,尽量剔除掉,以免干扰回归。用影响强度向量构造的权重向量 $\hat{\omega}$ 替代惩罚项的权重向量 ω 就得到了加权 Lasso 惩罚项。

下面给出同期解释变量复杂网络的构建和权重向量的计算方法。记 ADL 模型 2 解释变量的所有 p 阶滞后为 $x^p_t = (z^2_{1,t-p}, \cdots, z^2_{N_2,t-p})$, 其中, $p = 1, \cdots, P$, 把 x^p_t 中的所有元素作为网络的节点,共 N_2 个,构建 p 阶滞后复杂网络的基本步骤[13]如下:

1. 用 ρ_{ij} 表示第 i 个节点 $x^p_{i,t}$ 与第 j 个节点 $x^p_{j,t}$ 的 Pearson 相关系数,并计算两个节点的相似性: $S(i,j) = |\rho_{ij}|$, 得到相似矩阵 S, 显然, $S(i,j) \in [0,1]$。

2. 引入参数 b, 对 S 进行如下变化: $a_{ij} = S^b(i,j)$, 得到邻接矩阵 a, a 展示了网络中两个节点的连接方式。对于节点 $x^p_{j,t}(j = 1, \cdots, N_2)$, 定义影响强度: $k_j = \sum_{u \neq j} a_{j,u}, u = 1, \cdots, N_2$。记影响强度 $k = (k_1, \cdots, k_{N_2})$ 的分布为 $p(k)$。若 $p(k)$ 服从幂律,即:

$$p(k) \sim k^{-\gamma} \quad (5)$$

其中 γ 是幂律指数,则此网络具有无标度拓扑特性[14]。那么,如何量化无标度拓扑特征呢?事实上,若用 $\log(k)$ 拟合 $\log(p(k))$, 可用其拟合优度 R^2 来衡量无标度拓扑特征, R^2 越接近于 1, 无标度拓扑特征越明显。

不同的参数 b 对应着不同的网络构造方法,考虑到无标度拓扑特征可以凸显重要的节点,所以我们需要选择合适的 b, 使得构造的网络具有尽量明显的无标度拓扑特征。注意到, b 是 $S^b(i,j)$ 的幂指数,倘若相似矩阵中的元素 $S(i,j) \in (0,1)$, 那么不论 b 怎么变化,都不会把相似矩阵中的值强硬地变换成 0 或者 1, 我们把这个参数称为软阈值。

但如果软阈值取值过大,会使得 $S^b(i,j)$ 接近于 0, 即 a_{ij} 接近于 0, 这样 k^p 也会非常小,从而影响后续利用 k^p 构造出的权重向量的有效性。所以,我们在选择软阈值时,会先将其限制在一定的范围内(本文限制软阈值在 1 ~ 20 内),然后选择此取值范围内能够使得网络无标度拓扑特征最明显的幂指数。已知可以用 R^2 来量化无标度拓扑特征,那么只需计算出给定取值范围内所有幂指数构造出的网络所对应的 R^2, 然后选择使得 R^2 最大的软阈值,并计算此时的影响强度,记为 k^p。

用同样的方法计算出所有解释变量的影响强度并计算权重向量:

$$\hat{\omega} = (1/k^1_1, \cdots, 1/k^p_1, \cdots, 1/k^1_{N_2}, \cdots 1/k^p_{N_2})$$

网络结构 ADL 模型同样需要借助 LARS 算法实现,网络结构 ADL 模型的实现方法如下:

1. 计算网络结构权重向量 $\hat{\omega}$;

2. 定义: $x^*_j = x_j / \hat{\omega}_j, j = 1, 2, \cdots, N$;

3. 用 LARS 算法求解以下 Lasso 模型:

$$\hat{\beta}^* = \underset{\beta}{\operatorname{argmin}} \{ \| Y - \sum_{j=1}^{N} x^*_j \beta_j \|^2 + \lambda \sum_{j=1}^{N} |\beta_j| \}$$

4. 输出最终的参数估计量和变量选择结果:

$$\hat{\beta}_j = \hat{\beta}^* / \hat{\omega}_j, j = 1,2,\cdots,N$$

(四)Lasso-ADL 模型 3

令:

$$x_t = (y_{t-1},\cdots,y_{t-P},z^1_{1,t-1},\cdots,z^1_{1,t-P},\cdots,z^1_{N_1,t-1},\cdots,z^1_{N_1,t-P},z^2_{1,t-1},\cdots,z^2_{1,t-P},\cdots,z^2_{N_2,t-1},\cdots z^2_{N_2,t-P})$$

则此时的 ADL 模型可写成:

$$y_t = \alpha_0 + \beta' x_t + \varepsilon_t = \alpha_0 + \sum_{p=1}^{P}\alpha_p y_{t-p} + \sum_{j=1}^{N_1}\sum_{p=1}^{P}\varphi^1_{jp} z^1_{j,t-p} + \sum_{l=1}^{N_2}\sum_{p=1}^{P}\varphi^2_{lp} z^2_{l,t-p} + \varepsilon_t \quad (6)$$

本模型将被解释变量的自身信息、结构化数据和非结构化数据一起放入模型,这是较常用的处理方法,考察此模型的目的是为了与区别对待两类信息的 ADL 模型 2 进行对比,本文将此模型简称为"ADL 模型 3",此处将采用 ADL 模型 3 作为基本预测模型,同样使用 Lasso 对此模型进行变量选择。

三、数据来源及数据预处理

(一)数据来源

样本数据为深圳、上海、北京、广东、天津、湖北、重庆 7 个试点从启动日至 2016 年 5 月 10 日的每日碳价。其中,深圳碳交易试点启动最早,其样本数据为 2013 年 6 月 18 日至 2016 年 5 月 10 日的每日碳价,一共 640 组数据。而重庆碳交易试点只有 29 组数据,且间距较大,本文不对重庆试点进行预测研究。

结构化数据为与中国碳价密切相关的 5 个官方统计指标,样本为 2013 年 6 月 18 日至 2016 年 5 月 10 日的日频数据。

其中,*Ecarbix* 为欧洲碳指数现货每日结算价,取自欧洲能源交易所(EEX)。*EUAF* 为欧盟碳排放配额(EUA)连续期货合约日结算价,*CERF* 为核证减排量(CER)连续期货合约日结算价,均取自 Wind 数据库。*EREU* 和 *ERD* 为汇率,取自国家外汇管理局。

非结构化信息采用了百度整体搜索指数和媒体指数。百度搜索指数是以网民在百度的搜索量为数据基础,以关键词为统计对象,计算出的每个关键词在百度网页搜索中搜索频次的加权和。百度媒体指数是以各大互联网媒体报道的新闻中,与关键词相关的、被百度新闻频道收录的数量,采用新闻标题包含关键词的统计标准,与百度搜索指数无直接关系,两者均可在百度指数网站获得。虽然经过标准化处理,但本质上是非结构化的,属于非结构化数据[8]。

本文首先借助百度需求图谱的需求分布及相关词分类①对与碳排放权价格紧密关联的关键词进行初步选取,然后通过多轮小组讨论和专家审核的方式对初选的关键词进行筛选和增补,最后选定碳排放、低碳、环保、口罩、雾霾等 34 个最有可能影响碳价的关键词,搜集其 2013 年 6 月 18 日至 2016 年 5 月 10 日每日的百度搜索指数和媒体指数,用来衡量网民和互联网媒体对碳价的关注程度②,见表 1。

表 1　　百度指数

变量	含义	关键词
$BS_1,\cdots,BS_{34}$	百度搜索指数	北京雾霾、低碳和低碳经济等
$BM_1,\cdots,BM_8$	百度媒体指数	环保、口罩和生态文明等

注:34 个关键词中只有 8 个拥有较完整的百度媒体指数,所以媒体指数变量的个数为 8。

(二)数据预处理

虽然深圳碳排放交易试点碳排放权均价、官方统计指标及百度指数样本数据的起止时间相同,但受国内外公共节假日和交易时间不同等因素的影响,三者的样本容量并不相同,所以,在对所有数据进行平稳性变换之后③,将选择公共样本时间段作为研究区间。

经过预处理后的深圳试点样本数据一共有 543 组,按照时间顺序对数据集进行分割,后 25 个数据为预测集,剩下的为训练集,上海、北京、广东、天津、湖北这 5 个试点做同样处理。本文所有结果均由 R 软件呈现。

四、实证结果

为了防止过度拟合,当样本容量不是很高的时候,滞后阶数不宜过高,限定 P 为 2。

(一)实证结果分析

使用第二部分列出的 4 个模型对 6 个试点的碳价分别进行拟合和预测,由于 6 个试点使用 Lasso-

① 需求分布是综合计算关键词与相关词的相关程度,以及相关词自身搜索需求的大小得出。相关词分类分为来源相关词、去向相关词、搜索指数和上升最快。来源相关词是指网民在搜索关键词之前的搜索需求,去向相关词是指搜索关键词之后的搜索需求,搜索指数是指关键词的所有相关词中搜索指数最高的排名,上升最快是指所有关键词中搜索指数上升速度的排名。

② 有需要百度指数关键词的完整数据的读者可向作者索取。

③ 平稳性变换公式为:$\ln(x_t/X_{t-1})$。

ADL 模型 3 选择的最优模型中未选入任何变量，所以表 2 只列出了前 3 个模型的详细实证结果。

表 2 前 3 个模型的详细实证结果(%)

模型	试点	模型变量	训练集 MSE	预测集 MSE
Lasso-ADL 模型 1	深圳	*ySZA*1、*CERF*1	0.9577	1.8477
	上海	0	0.3538	0.4266
	北京	*yBEA*1、*CERF*2	0.3302	1.2422
	广东	*yGDEA*1、*yGDEA*2、*Ecarbix*1、*EUAF*1、*EUAF*2、*CERF*1、*CERF*2	0.4916	0.8454
	天津	*yTJEA*1、*EUAF*2、*CERF*1、*CERF*2	0.1524	0.0026
	湖北	0	0.0591	0.1855
Lasso-ADL 模型 2	深圳	*ySZA*1、*CERF*1	0.9577	1.8477
	上海	0	0.3538	0.4266
	北京	*yBEA*1、*CERF*2	0.3302	1.2422
	广东	*yGDEA*1、*yGDEA*2、*Ecarbix*1、*EUAF*1、*EUAF*2、*CERF*1、*CERF*2	0.4916	0.8454
	天津	*yTJEA*1、*EUAF*2、*CERF*1、*CERF*2	0.1524	0.0026
	湖北	$BS_3$1(低碳经济)、$BM_1$1(环保)	0.0560	0.1762
网络结构 ADL 模型	深圳	*ySZA*1、*CERF*1	0.9577	1.8477
	上海	0	0.3538	0.4266
	北京	*yBEA*1、*CERF*2、BS_{31}1(碳交易)	0.3257	1.2394
	广东	*yGDEA*1、*yGDEA*2、*Ecarbix*1、*EUAF*1、*EUAF*2、*CERF*1、*CERF*2	0.4916	0.8454
	天津	*yTJEA*1、*EUAF*2、*CERF*1、*CERF*2	0.1524	0.0026
	湖北	$BS_2$2(低碳)、BS_{23}1(节能)、BS_{31}1(碳交易)、$BM_1$1(环保)	0.0549	0.1639

注：变量尾数 1,2,…分别表示滞后 1 期，滞后 2 期，…，比如，*ySZA*1 表示深圳试点滞后 1 期的碳价。对于百度指数，括号里给出了变量对应的关键词。

表 2 中 Lasso-ADL 模型 1 的解释变量为试点碳价的自身滞后和官方统计指标，采用 Lasso 进行模型选择。Lasso-ADL 模型 2 是在 Lasso-ADL 模型 1 的基础上对应地增加百度指数(百度整体搜索指数和百度媒体指数)后得到的结果，主要考察非结构信息对预测的影响。网络结构 ADL 是对 Lasso-ADL 模型 2 的改进，是为了在回归时兼顾解释变量之间的网络关系而提出的，可以提高预测效果。

为了从总体上对比分析不同模型的预测效果，计算 6 个试点训练集 MSE 的平均，作为模型的训练集 MSE 均值，同样的方法产生预测集 MSE 均值，详见表 3。

表 3 所有模型的实证结果 (%)

模型	训练集 MSE 均值	预测集 MSE 均值
Lasso-ADL 模型 1	0.3908	0.7583
Lasso-ADL 模型 2	0.3903	0.7568
网络结构 ADL 模型	0.3894	0.7552
Lasso-ADL 模型 3	0.4219	0.8398

对表 2 和表 3 进行分析，得出如下结论：

1. 非结构化信息有助于提高碳价预测能力。

从总体上看，Lasso-ADL 模型 2 训练集 MSE 均值和预测集 MSE 均值都低于 Lasso-ADL 模型 1，网络结构 ADL 模型同样如此。Lasso-ADL 模型 2 和网络结构 ADL 模型都使用了非结构化信息，而 Lasso-ADL 模型 1 并没有使用。所以，加入非结构化信息可以改善预测效果。

从细节上来看，Lasso-ADL 模型 2 训练集和预测集 MSE 均低于 Lasso-ADL 模型 1 的关键在于湖北试点。湖北试点在没有考虑非结构化信息时最优模型内没有选入任何变量，而在考虑了非结构化信息之后发现，关键词“低碳经济”的百度整体搜索指数变量 1 阶滞后和关键词“环保”的百度媒体指数变量 1 阶滞后的加入大大减小了训练集和预测集的 MSE。同样的方法对比分析网络结构 ADL 模型与 Lasso-ADL 模型 1，发现前者优于后者的原因是北京和湖北试点的最优模型中加入了搜索指数和媒体指数变量，非结构化信息提高了碳价的预测效果。

2. 只有选择了合适的预测模型，大数据的价值才能被体现。

Lasso-ADL 模型 2、网络结构 ADL 模型和 Lasso-ADL 模型 3 的解释变量是相同的，均包括了被解释变量的自身信息、结构化信息和非结构化信息，但是预测效果却不同。Lasso-ADL 模型 3 平等对待所有变量，导致没有一个变量被选入最终模型，从而取得了非常差的预测效果，甚至不如没有考虑非结构信息的 Lasso-ADL 模型 1。Lasso-ADL 模型 2 和网络结构 ADL 模型都是先用被解释变量自身信息和结构化信息拟合被解释变量，再加入非结构化信息，大幅提高了预测效果，较好地挖掘了大数据的价值。所以，选择合适的预测模型是使用大数据进行预测的关键。

3. 网络结构 ADL 模型更适用于基于大数据的碳价预测。

Lasso-ADL 模型 2、网络结构 ADL 模型和 Lasso-ADL 模型 3 都是基于大数据的预测模型。

从总体上看，网络结构 ADL 模型在所有模型中表现最好，因为其训练集 MSE 均值和预测集 MSE 均值最小。

从细节上来看，Lasso-ADL 模型 3 的最优模型中没有选入任何变量，是所有模型中表现最差的。Lasso-ADL 模型 2 使湖北试点选了 2 个百度指数变量，提高了湖北试点的预测能力，优于 Lasso-ADL 模型 3。不过，此模型没有考虑变量之间的网络关系，

网络结构 ADL 模型在考虑了网络结构之后,不仅使湖北试点的最优模型中选入了更合适的百度整体指数和媒体指数变量,进一步提高了预测效果,而且为北京试点的碳价预测最优模型也选入了一个合适的百度整体指数变量,提高了此试点的预测效果。所以,网络结构 ADL 模型更适合基于大数据的预测。

(二)对比分析

从实证结果中可以看出,网络结构 ADL 模型最适合基于大数据的碳价预测。为了进一步检验网络结构 ADL 模型的有效性,本部分将使用滚动窗口对碳价进行预测。

以时间为序,选择后 60 个数据为滚动预测集,每个滚动窗口的预测集样本容量仍旧为 25,所以,每个试点可以得到 35 个滚动窗口。每个滚动窗口都可以计算出一个训练集 MSE 均值和预测集 MSE 均值,对所有滚动窗口的训练集 MSE 均值进行平均,得到滚动训练集 MSE 均值,同理计算滚动预测集 MSE 均值。计算结果见表 4。

表 4　　滚动窗口实证结果　　(%)

模型	滚动训练集 MSE 均值	滚动预测集 MSE 均值
Lasso-ADL 模型 2	0.3520	1.0461
网络结构 ADL 模型	0.3519	1.0452
Lasso-ADL 模型 3	0.3794	1.1288

若用 MSE_{net} 来表示使用网络结构 ADL 模型得到的 MSE,MSE_2 表示使用 Lasso-ADL 模型 2 得到的 MSE,MSE_3 表示使用 Lasso-ADL 模型 3 得到的 MSE,对每个滚动窗口计算如下比率:

$$ratio_2 = \frac{MSE_{net}}{MSE_2} \tag{7}$$

$$ratio_3 = \frac{MSE_{net}}{MSE_3} \tag{8}$$

图 1 是所有滚动窗口训练集和预测集 $ratio_2$ 的直方图,图 2 是所有滚动窗口训练集和预测集 $ratio_3$ 的直方图。

接下来,利用表 4、图 1 和图 2 对 3 个基于大数据的预测模型进行对比分析。

1. 网络结构 ADL 模型与 Lasso-ADL 模型 2 的比较分析。

表 4 中网络结构 ADL 模型的滚动训练集 MSE 均值和滚动预测集 MSE 均值都小于 Lasso-ADL 模型 2,说明总体上,网络结构 ADL 模型更优。

图 1 训练集 $ratio_2$ 出现小于 1 的次数多于出现大于 1 的次数。预测集 $ratio_2$ 出现小于 1 的次数也

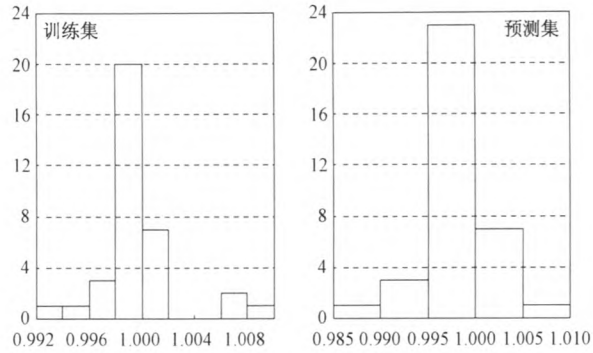

图 1　滚动窗口训练集和预测集 $ratio_2$ 的直方图

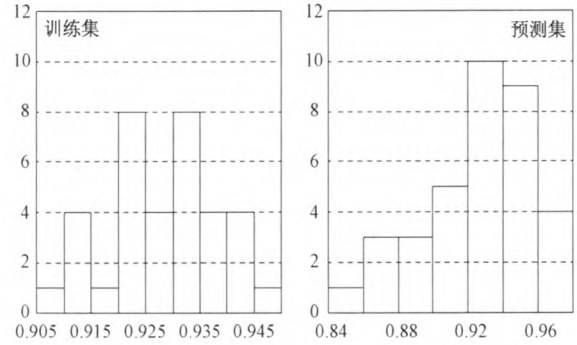

图 2　滚动窗口训练集和预测集 $ratio_3$ 的直方图

明显多于出现大于 1 的次数,说明网络结构 ADL 仍旧表现更好。

所以,考虑变量之间的网络结构的确提高了模型的预测准确性。

2. 网络结构 ADL 模型与 Lasso-ADL 模型 3 的比较分析。

Lasso-ADL 模型 3 不区别对待任何解释变量,仅利用 Lasso 模型对所有解释变量进行选择,而网络结构 ADL 模型不仅区别对待了结构化信息和非结构化信息,而且还兼顾了变量之间的网络结构。从表 4 中可以看出,网络结构 ADL 模型的滚动训练集 MSE 均值和滚动预测集 MSE 均值都远远低于 Lasso-ADL 模型 3,从图 2 中可以看出,所有的 $ratio_3$ 都小于 1,即网络结构 ADL 模型显著地优于 Lasso-ADL 模型 3。

经过以上分析可知,在使用大数据进行预测时,网络结构 ADL 是最优的。

五、结论

大数据为预测提供了更多的资源,但是,数据并不是越多越好,如何有效利用大数据进行碳价预测是本文研究的主要内容。事实上,基于大数据的预

测模型中的解释变量往往非常多,考虑到解释变量之间的网络结构会影响碳价预测的准确性,本文提出了网络结构 ADL 模型,显著地提高了预测效果。

本文的主要工作和结论如下:第一,为了明确大数据是否有助于提高预测效果,本文同时使用不含非结构化信息的预测模型和加入了非结构化信息的预测模型对碳价进行预测,得出非结构化信息有助于提高碳价预测效果的结论。第二,为了明确预测模型是否会影响基于大数据的碳价预测效果,本文使用相同的解释变量和不同的预测模型对碳价进行预测,发现平等对待所有解释变量的 Lasso-ADL 模型 3 表现最差,甚至不如没有考虑非结构化数据的 Lasso-ADL 模型 1,只有选择了正确的预测模型才能充分发挥大数据价值。第三,为了考察本文提出的网络结构 ADL 模型的优越性,对 Lasso-ADL 模型 2、网络结构 ADL 模型和 Lasso-ADL 模型 3 进行对比分析。分析发现,网络结构 ADL 模型可以挑选更有效的百度指数变量,显著地提高碳价预测能力。使用滚动窗口进一步检验新模型的有效性,同样可发现网络结构 ADL 模型是使用大数据进行碳价预测的最优模型,为使用大数据进行预测提供了一种有效且稳定的新方法。

参考文献

[1] B Z Zhu, Y M Wei. Carbon price forecasting with a novel hybrid ARIMA and least squares support vector machines methodology [J]. Omega, 2013(41): 517 - 524.

[2] W Li, C Lu. The Research on Setting a Unified Interval of Carbon Price Benchmark in the National Carbon Trading Market of China [J]. Applied Energy, 2015(155): 728 - 739.

[3] N A Armah. Big data analysis: the next frontier[J]. Bank of Canada Review, 2013(Summer): 32 - 39.

[4] H Choi, H Varian. Predicting initial claims for unemployment benefits[EB/OL]. http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/fr//archive/papers/initialclaimsUS.pdf, 2009 - 04 - 10.

[5] N Askitas, K F Zimmermann. Google econometrics and unemployment forecasting [J]. Applied Economics Quarterly, 2009(2): 107 - 120.

[6] H Choi, H Varian. Predicting the present with google trends[J]. Economic Record, 2012, 88(supplement s2): 2 - 9.

[7] G L Schlegel. Utilizing big data and predictive analytics to manage supply chain risk [J]. Journal of Business Forecasting, 2014(4): 11 - 17.

[8] 刘涛雄, 徐晓飞. 互联网搜索行为能帮助我们预测宏观经济吗? [J]. 经济研究, 2015(12): 68 - 83.

[9] J Huang, S G Ma, H Z Li, et al. The sparse Laplacian shrinkage estimator for high-dimensional regression [J]. The Annals of Statistics, 2011(4): 2021 - 2046.

[10] 方匡南, 范新妍, 马双鸽. 基于网络结构 Logistic 模型的企业信用风险预警[J]. 统计研究, 2016(4): 50 - 55.

[11] R Tibshirani. Regression Shrinkage and Selection via the Lasso[J]. Journal of the Royal Statistical Society, Series B, 1996(1): 267 - 288.

[12] B Efron, T Hastie, I Johnstone, et al. Least Angle Regression [J]. Annals of Statistics, 2004 (2): 407 - 499.

[13] B Zhang, S Horvath. A General Framework for Weighted Gene Co-Expression Network Analysis [J]. Statistical Applications in Genetics and Molecular Biology, 2005, 4(1): 1 - 37.

[14] AL Barabási, R Albert. Emergence of scaling in random network [J]. Science, 1999, 286(5439): 509 - 512.

作者简介

王娜,女,2012 年毕业于成都理工大学,获计算数学专业理学硕士学位,现为厦门大学经济学院统计学在读博士研究生。研究方向为经济统计。

(责任编辑:倪立行)