

基于非结构数据流行学习的碳价格多尺度组合预测

刘金培^{1,2†}, 郭 艺^{1,3}, 陈华友⁴, 任贺松^{1,5}, 陶志富⁶

(1. 安徽大学 商学院, 合肥 230601; 2. 北卡罗来纳州立大学 工业与系统工程系, 美国 罗利 27695;
3. 东南大学 经济管理学院, 南京 211189; 4. 安徽大学 数学科学学院, 合肥 230601;
5. 西安交通大学 管理学院, 西安 710049; 6. 安徽大学 经济学院, 合肥 230601)

摘 要: 碳交易价格的有效预测对制定符合国情的碳金融市场政策以及碳金融市场的风险管理都具有重要意义. 对此, 提出一种基于非结构数据流行学习的碳价格多尺度组合预测方法. 首先, 利用网络搜索指数提取碳价格相关的非结构化数据, 基于等度量映射流行学习对其进行降维; 然后, 对降维后的非结构化数据、其他影响因素结构化数据、碳交易价格分别进行经验模态分解 (Empirical mode decomposition, EMD), 得到不同个数的本征模函数 (Intrinsic mode function, IMF), 并采用 Fine-to-coarse 方法对 IMF 进行重构, 得到高频序列、低频序列和趋势项; 最后, 利用自回归积分滑动平均模型 (Autoregressive integrated moving average model, ARIMA)、偏最小二乘 (Partial least squares, PLS) 回归和神经网络对高频数据、低频数据和趋势项进行预测, 将 3 种预测结果进行集成, 得到最终预测值. 仿真实验结果表明, 所提出的方法可以有效利用多源信息, 具有较高的预测精度和良好的适用性.

关键词: 多尺度预测; 碳价格; 非结构数据; 流行学习; 自回归积分滑动平均模型; 经验模态分解

中图分类号: O212

文献标志码: A

Multi-scale combined forecast of carbon price based on manifold learning of unstructured data

LIU Jin-pei^{1,2†}, GUO Yi^{1,3}, CHEN Hua-you⁴, REN He-song^{1,5}, TAO Zhi-fu⁶

(1. School of Business, Anhui University, Hefei 230601, China; 2. Department of Industrial and Systems Engineering, North Carolina State University, Raleigh 27695, USA; 3. School of Economics and Management, Southeast University, Nanjing 211189, China; 4. School of Mathematical Sciences, Anhui University, Hefei 230601, China; 5. School of Management, Xi'an Jiaotong University, Xi'an 710049, China; 6. School of Economics, Anhui University, Hefei 230601, China)

Abstract: Forecasting carbon trading price effectively is of great significance for the formulation of carbon financial market policies that suit to Chinese condition and the risk management of carbon financial market. This paper proposes a multi-scale combined forecast method for carbon price based on manifold learning of unstructured data. Firstly, the unstructured data related to carbon price is extracted using the network search index, and dimensionality reduction is performed based on the isometric mapping manifold learning. Then, the structured data of other influencing factors and the carbon trading price are decomposed into a variable number of intrinsic mode functions (IMFs) using empirical mode decomposition (EMD) respectively. The IMF is reconstructed to get high frequency sequence, low frequency sequence and trend item based on the fine-to-coarse method. Moreover, autoregressive integrated moving average model (ARIMA), Partial least squares (PLS) regression and neural network are used to forecast high-frequency data, low-frequency data and trend items, which are aggregated to get the final forecast result. The results of simulated experiments show that the proposed method can effectively use multi-source information and has high prediction accuracy and good applicability.

Keywords: multi-scale forecasting; carbon price; unstructured data; manifold learning; ARIMA; EMD

收稿日期: 2018-01-26; 修回日期: 2018-08-21.

基金项目: 国家自然科学基金项目 (71501002, 61502003, 71771001, 71701001); 安徽省自然科学基金项目 (1608085 QF133, 1508085QG149); 安徽省高校省级自然科学研究重点项目 (KJ2017A026).

责任编委: 李登峰.

作者简介: 刘金培 (1984—), 男, 副教授, 博士, 从事预测与决策分析等研究; 郭艺 (1996—), 女, 硕士生, 从事组合预测的研究.

†通讯作者. E-mail: liujinpei2012@163.com.

0 引言

近年来,我国环境污染问题频发,碳排放备受人们的关注.作为全球最大温室气体排放国,我国一直致力于对碳排放的控制,“十三五”规划指出,主动控制碳排放量,落实减排承诺,推动建设全国统一的碳排放交易市场.我国已启动上海、深圳等7个碳排放权交易试点,以发电行业为突破口的全国碳交易体系于2017年12月正式启动.碳交易过程的核心因素之一是碳交易价格,因此,有效预测碳交易价格既能帮助投资者进行科学决策,也有利于政府制定合理的碳金融交易政策,对引导碳排放市场的科学理性发展具有重要意义.目前,碳价格预测已成为国内外学者研究的热点.最初对碳价格进行预测主要是基于经典的统计和计量模型,如AR(1)-GARCH(1,1)域^[1]、非参数统计域^[2]、GED-GARCH域^[3]、灰色马尔科夫域^[4]等预测方法.然而,由于碳交易价格时间序列具有高度的非线性和非平稳性,传统的统计和计量模型往往无法有效处理这种非线性模式.此后,结合机器学习和人工智能的预测技术得到了快速发展,如ARIMA-SVM域^[5]、基于多层感知器的神经网络模型域^[6]、 k 近邻算法域^[7]等预测模型.

最近研究表明,将时间序列进行多尺度分解,分解后的序列具有更好的尺度波动规律性,再选择恰当的模型分别预测分解得到的序列,具有更高的预测精度^[8-12].小波分解^[8]和经验模态分解(Empirical mode decomposition, EMD)^[9]是分解多尺度时间序列的常用方法.与小波分解方法不同,EMD不需要预先设定分解尺度,因此自适应性更强^[10],可以完整分解非平稳时间序列信号,具有更高的拟合精度.Zhu等^[11]首先将碳价格时间序列进行EMD分解,然后利用改进的最小二乘支持向量回归对分解后的时间序列分别进行预测,取得了较好的预测效果.Li等^[12]则提出在将碳价格时间序列进行EMD分解的基础上,利用广义自回归条件异方差(GARCH)模型对分解后的序列进行预测,并对上海、深圳等5个试点的碳交易价格进行了预测.高杨等^[13]提出了基于EMD分解的粒子群算法(PSO)-支持向量机(SVM)误差校正预测模型,对碳金融国际市场价格进行预测,预测精度明显提高.实际上,利用EMD将时间序列分解后,各本征模态函数(IMF)序列对应的尺度波动特征不同^[14],若采取同一种方法对分解后的高频和低频序列进行预测,则往往会导致适应度较差.王书平等^[10]将时间序列数据多尺度分解后,使用不同的方法预测不同特点的分量,实证结果表明,采用不同方法预测的精度高于

其他基于EMD的多尺度预测模型.

碳交易价格预测除了需考虑市场环境和能源价格等结构化数据影响因素外^[15-16],基于网络搜索信息的非结构化数据也为碳价格预测提供了大量的有效信息^[17].最新研究表明,将网络搜索指数非结构化数据应用到流感趋势、房地产价格、股票价格等领域预测中均明显提升了预测的准确性^[18].董倩等^[19]在拟合并预测16个城市的二手房和新房价格时以百度指数为基础,结果表明,在准确预测房价指数以及反映经济主体行为规律与趋势时,网络搜索数据发挥了显著作用.王炼等^[20]以消费者网络搜索信息为基础,利用fractional logit模型验证了市场份额受消费者网络搜索的影响.陈植元等^[21]将百度指数应用于股票市场的研究,采用固定效应常系数面板数据模型对搜索数据进行分析,有效改善了已有股票预测模型的准确度.颜七笙等^[22]针对国际铀资源价格预测问题,提出了一种基于经验模式分解(EMD)、相空间重构(PSR)和极限学习机(ELM)的非线性组合预测方法.然而,基于非结构化信息的碳价格非线性预测方法的文献还比较少见.

从已有研究来看,关于碳交易价格预测还存在以下两个问题:1)虽然基于网络搜索指数的非结构化数据为碳价格预测提供了大量的有效信息,然而如何利用这些信息进行碳排放交易价格预测,尤其是基于非结构化数据的碳价格多尺度预测,还缺乏深入的研究;2)已有基于EMD的碳价格预测方法大多仅利用了碳价格历史数据,而考虑到能源价格、市场环境等相关因素影响的文献很少.能够有效提取来自各维度的有效信息是组合预测能够提高预测精度的原因^[23].目前,尚未见到综合考虑非结构化和结构化数据的碳价格多尺度组合预测相关研究.

为了综合利用多源信息,考虑到碳价格并非平稳和非线性的特点,本文提出一种基于非结构数据流行学习的碳价格多尺度组合预测方法.首先,从百度搜索指数提取碳价格相关的非结构化数据,基于等度量映射流行学习对其进行降维处理;然后,把降维后的非结构化数据、其他影响因素结构化数据和碳价格分别使用EMD分解,并对分解得到的IMF进行重构,得到高、低频率序列和趋势项;进而针对重构后数据序列的不同特点,用ARIMA、偏最小二乘(PLS)回归和神经网络对高频数据、低频数据和趋势项进行预测,将3组时间序列预测的结果集成,得到最终价格的预测值;最后,通过仿真实验来检验该组合预测模型的准确性与适用性.

1 多尺度组合预测方法

1.1 碳市场交易价格的影响因素

国内碳市场交易价格的波动受到供给、需求等多方因素的影响. 首先,碳排放配额是碳市场交易价格最大的供给方面影响因素^[24],同时,欧盟碳排放配额(EUA)价格是全球碳交易市场的基准,对碳排放配额有重要影响. 考虑到市场的联动性,本文选取欧盟碳排放配额价格作为国内碳市场价格的影响因素. 其次,碳排放与化石能源使用密切相关,而化石能源的需求决定于它们的绝对和相对价格^[24]. 焦煤指数、燃油指数、能源指数均为与碳排放量密切相关的经济指数,因此,本文用这3个指数来反映化石能源对碳市场交易价格的影响. 此外,碳交易价格还受到市场中其他因素的影响,本文选取上证B股指数来反映我国股票市场交易的情况,同时考虑到人民币对美元汇率对国内碳市场价格的影响.

除了上述结构化的影响因素数据可以反映碳价格波动外,网络中获取的非结构化数据也具有重要的补充作用. 随着互联网和大数据技术的快速发展,网络搜索大数据已成为反映投资者行为以及其对市场关注程度的重要指标. 目前,百度指数和谷歌趋势提供网络搜索数据服务,百度搜索在国内使用范围广,具有更高的可靠性^[21],所以本文根据百度需求图谱和相关词分类,选取碳足迹、低碳经济、碳交易、碳排放、碳汇、减排、碳中和等密切相关的关键词的百度搜索指数.

1.2 多尺度组合预测方法基本框架

碳市场交易价格时间序列受到市场环境和能源价格等不稳定因素的影响,呈现出非线性、非平稳和多频率叠加等特征. 本文提出多尺度组合预测方法框架如图1所示.

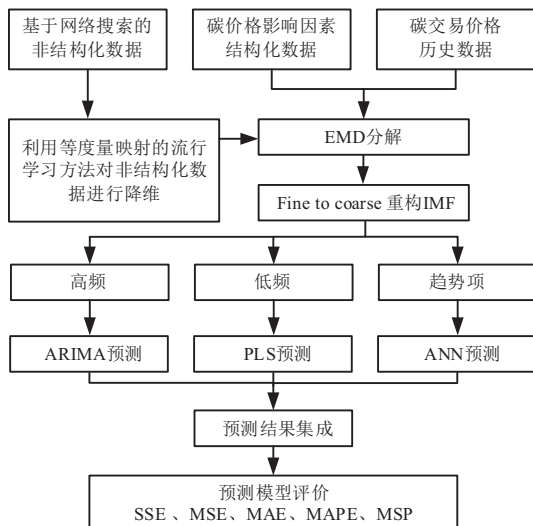


图1 碳价格多尺度组合预测框架

首先,利用等度量映射的流行学习方法对非结构化数据进行降维,再把降维后的非结构化数据、结构化数据和碳价格数据分别用EMD分解为多个具有不同频率的IMF,并运用Fine-to-coarse方法把EMD分解的数据重构为高频、低频、趋势项;然后,选择适合的预测方法来预测不同的频率分量;最后,将各频率分量的预测结果进行集成. 为了验证多尺度组合预测模型的精度,选用误差平方和(SSE)、平均绝对百分比误差(MAPE)、平均绝对误差(MAE)、均方误差(MSE)和均方百分比误差(MSPE)这5种常见的指标对预测结果进行检验和评价. 与已有碳价格预测方法相比,本文建立的碳价格多尺度组合预测方法具有3个优势: 1)充分利用非结构化信息,并且对非结构化信息进行基于等度量映射流行学习的降维处理,该降维方法可以保持近邻样本间的距离不变; 2)将碳价格数据、影响因素数据和流行学习后的非结构化信息分别进行多尺度分解,提取不同的频率特征,在充分利用多源信息的基础上,可进一步提升预测精度; 3)针对高频、低频和趋势项的特点,分别选取不同的单项预测模型,最后将预测结果进行集成,具有较好的自适应性.

1.3 多尺度组合预测模型构建的基本过程

1.3.1 基于流行学习的非结构化数据降维

基于网络搜索的非结构化信息不但具有很高的维数,而且具有一定的信息冗余. 为了降低信息冗余,有效利用非结构化数据,本文基于流行学习方法对其进行降维处理. 非线性降维在流行学习基础上,与传统降维相比能够在局部保持欧氏空间的性质. 设碳交易价格时间序列为 $y_t, t = 1, 2, \dots, n$,若在 t 时刻与碳交易价格相关的 m 维非结构化样本数据为 $x_t = (x_{t1}, x_{t2}, \dots, x_{tm})^T$,则非结构化数据矩阵为 $X = (x_{ij})_{n \times m}$. 为了将 X 降为 d 维的数据矩阵 $Z = (z_{ij})_{n \times d}$,基于等度量映射(Isomap)流行学习的降维步骤如下^[25].

Step 1: 对于样本数据中任意两个时刻数据点 x_i 与 x_j ,计算其欧氏距离,即

$$d_{ij} = d(x_i, x_j) = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2},$$

$$i, j = 1, 2, \dots, n. \tag{1}$$

Step 2: 对于样本数据 x_i ,确定其 ϵ -领域内的点组成的集合 A_i . 以样本数据为点建立加权图, x_i 与 x_j 之间的边权值由下式求得:

$$w_{ij} = w_{ij}(x_i, x_j) = \begin{cases} d_{ij}, & x_j \in A_i; \\ \infty, & x_j \notin A_i. \end{cases}$$

Step 3: 对于加权图内任意两节点, 根据 Floyd 算法计算其最短距离, 记为 d'_{ij} . 令 $d'_i = \frac{1}{n} \sum_{j=1}^n d'_{ij}$, $d'_{..} = \frac{1}{n} \sum_{i=1}^n d'_i$, $d'_{..} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d'_{ij}$, 进而可得到降维后的内积矩阵 $B = ZZ^T = (b_{ij})_{n \times n}$, 其中

$$b_{ij} = -0.5(d'_{ij}{}^2 - d'_i{}^2 - d'_j{}^2 + d'_{..}{}^2). \quad (2)$$

Step 4: 对矩阵 B 进行特征值分解, 得到 $B = UAU^T$. 其中: A 为特征值构成的对角矩阵, 即 $A = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$, 且 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$; U 为特征值对应的特征向量矩阵. 取其中前 d 个特征值, 令 $A^* = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_d)$, U^* 为前 d 个特征值对应的特征向量矩阵.

Step 5: 得到降维后的数据矩阵 $Z = (A^*)^{0.5} \times (U^*)^T$, 此时 Z 中的每一行即为各样本降维后的数据.

由于碳市场价格序列具有非平稳、非线性的特点, 研究表明, 引入非结构化数据能使预测精度明显提高. 本文选取与碳交易价格密切相关的关键词的百度搜索指数作为碳交易价格相关的非结构化数据.

1.3.2 EMD分解

EMD 适用于具有非线性、非平稳特征的时间序列数据. 数据序列通过 EMD 筛选过程进行多尺度分解, 产生若干个本征模态函数 (Intrinsic mode function, IMF) 和一个趋势项. IMF 必须满足以下两个条件: 1) 过零点的数量与极值点的数量之差绝对值不大于 1; 2) 数据序列中任意一点对应的局部极小值点和局部极大值点确定的包络线均值为 0 且满足时间轴上的局部对称. EMD 分解的具体步骤如下:

Step 1: 识别出原时间序列 $x(t)$ 中局部的极小值点和极大值点.

Step 2: 运用三次样条插值处理所有极大值点和极小值点, 分别拟合出 $x(t)$ 的上包络线 $e_{\max}(t)$ 和下包络线 $e_{\min}(t)$.

Step 3: 原始时间序列 $x(t)$ 减去由上包络线与下包络线产生的均值序列 $m(t) = (e_{\max}(t) + e_{\min}(t))/2$, 得到去除低频的序列 $d(t) = x(t) - m(t)$.

Step 4: 用 $d(t)$ 代替原时间序列 $x(t)$, 重复 Step 1 ~ Step 3, 直到 $d(t)$ 满足 IMF 的定义. 此时, $d(t)$ 为第 1 个 IMF, 记为 $c_1(t)$, 令残差 $r(t) = x(t) - c_1(t)$ 为新的 $x(t)$.

Step 5: 重复 Step 1 ~ Step 4, 筛选出多个 IMF, 即 $c_i(t)$, 直到无法再从 $x(t)$ 中筛选出新的 IMF 为止. 通

过上述方法, 原时间序列 $x(t)$ 最终被分解为多个 IMF 与一个趋势项分量 $r(t)$, 即

$$x(t) = \sum_i c_i(t) + r(t).$$

1.3.3 Fine-to-coarse 重构

为了避免过度分解, 运用 Fine-to-coarse 技术把筛选的 IMF 重构成高频和低频两个部分. 若原时间序列 $x(t)$ 被分解为 N 个 IMF k , $k = 1, 2, \dots, N$, 则重构的具体步骤为:

Step 1: 令 $s_i = \sum_{k=1}^i \text{IMF}_k$, $i = 1, 2, \dots, N$.

Step 2: 利用 t 检验判断出 s_i 的均值显著不为 0 时所对应的 i 取值.

Step 3: 将 $\text{IMF}_1 \sim \text{IMF}_{i-1}$ 加总, 作为高频部分; 将 $\text{IMF}_{i-1} \sim \text{IMF}_N$ 加总, 作为低频部分.

1.3.4 单项预测模型

碳价格时间序列经 EMD 分解并 Fine-to-coarse 重构后, 根据各频率的不同特征, 分别选用 ARIMA、PLS 和神经网络对碳价格的高频 $y_1(t)$ 、低频 $y_2(t)$ 和趋势项 $y_3(t)$ 进行预测.

高频数据预测. 时间序列分析的重要工具之一是由自回归 (AR) 和移动平均 (MA) 组成的自回归移动平均模型 (ARIMA). 模型的公式一般如下:

$$y_1(t) = \varphi_1 y_1(t-1) + \varphi_2 y_1(t-2) + \dots + \varphi_p y_1(t-p) + \varepsilon(t) + \theta_1 \varepsilon(t-1) + \dots + \theta_q \varepsilon(t-q). \quad (3)$$

由于 ARIMA 能够对数据进行拆分使其平稳化, 利用该特点可以更加精确预测 EMD 分解后的高频数据, 得到碳价格高频预测值 $\hat{y}_1(t)$.

低频数据预测. PLS 回归结合了主成分分析、典型相关分析等多种方法, 可以克服非结构化数据与结构化数据相关性高的问题. 若非结构数据降维-分解-重构后在 t 时刻的低频数据记为 $z_2(t)$, 结构化影响因素数据分解重构后在 t 时刻的低频数据记为 $s_2(t)$, 则建立 $y_2(t)$ 与 $z_2(t)$ 、 $s_2(t)$ 之间的 PLS 回归模型, 即 $y_2(t) = f(z_2(t), s_2(t))$. 对所有数据首先进行标准化处理, 再提取第 1 成分, 进行回归分析, 若拟合度达标则建模完成; 否则进行迭代, 在剩余信息中实现第 2 次提取并回归. 如此反复直至拟合度达标, 根据回归方程得到碳价格低频预测值 $\hat{y}_2(t)$.

趋势项数据预测. 神经网络中连接各神经元的权重可通过反向传播误差算法求得. 用 BP 神经网络对趋势项数据进行预测时, 首先归一化处理数据, 并设置网络层数和每层的神经元数 n , 建立模型. 将非

结构数据降维-分解-重构后的趋势项 $z_3(t)$ 和结构化数据分解-重构的趋势项 $s_3(t)$ 作为输入, $y_3(t)$ 作为输出, 对神经网络进行训练. 将训练完成的神经网络对趋势项进行预测, 得到碳价格趋势项预测值 $\hat{y}_3(t)$.

1.3.5 预测结果的集成与评价

将 ARIMA 预测得到的碳价格高频预测值 $\hat{y}_1(t)$ 、PLS 回归预测得到的低频预测值 $\hat{y}_2(t)$ 和神经网络预测得到的趋势项预测值 $\hat{y}_3(t)$ 相加, 得到碳价格最终组合预测的预测值 $\hat{y}(t)$. 值得注意的是, EMD 分解后的高频序列、低频序列和趋势项在不同的频率空间里是独立的. 本文对 3 个序列预测值进行相加, 是原始 EMD 分解的逆过程, 可以保证预测结果的无偏性.

本文分别采用误差平方和 (SSE)、平均绝对百分比误差 (MAPE)、平均绝对误差 (MAE)、均方误差 (MSE) 以及均方百分比误差 (MSPE) 这 5 种常用的误差指标反映预测模型的预测精度^[8]. 误差指标的计算公式如下:

$$\begin{aligned} \text{MAE} &= \frac{1}{n} \sum_{t=1}^n |y(t) - \hat{y}(t)|, \\ \text{SSE} &= \sum_{t=1}^n (y(t) - \hat{y}(t))^2, \\ \text{MSE} &= \frac{1}{n} \sqrt{\sum_{t=1}^n (y(t) - \hat{y}(t))^2}, \\ \text{MAPE} &= \frac{1}{n} \sum_{t=1}^n \left| \frac{y(t) - \hat{y}(t)}{y(t)} \right|, \\ \text{MSPE} &= \frac{1}{n} \sqrt{\sum_{t=1}^n \left(\frac{y(t) - \hat{y}(t)}{y(t)} \right)^2}. \end{aligned}$$

2 仿真实验

2.1 数据来源及数据预处理

以上海市碳市场 2015 年 1 月 1 日 ~ 2016 年 12 月 31 日的碳价格数据为训练集, 以 2017 年 1 月 1 日 ~ 2017 年 3 月 10 日碳价格数据为测试集进行仿真实验分析.

非结构化数据来源为百度指数. 百度指数是一段时间内百度搜索数量中特定关键词的加权和, 再经过标准化处理得到的一类典型非结构化数据. 经过小组多轮讨论以及专家审核方式对初始的关键词进行筛选和补充, 结合百度需求图谱和相关词分类, 本文最后选定碳足迹、低碳经济、碳交易、碳排放、碳汇、减排、碳中和等 17 个最有可能影响碳交易价格的关键词, 利用这些关键词 2015 年 1 月 1 日 ~ 2017 年 3 月 10 日的百度搜索指数来衡量人们对碳交易的关注度.

碳价格影响因素的结构化数据, 选取与中国碳市场价格密切相关的 6 个统计指标: 欧盟碳排放配额 (EUA) 价格、焦煤指数、燃油指数、能源指数、上证 B 股指数和汇率. 其中: EUA 价格为欧盟碳排放配额 (EUA) 连续期货合约日结算价格, 来源于欧洲能源交易所 (EEX); 上证 B 股指数反映了中国股市交易的总体情况; 焦煤指数、燃油指数、能源指数均是与碳排放密切相关的经济指数, 来源于 Wind 数据库; 汇率数据来源于国家外汇管理局.

值得注意的是, 虽然碳平均排放权价格、百度指数以及各影响因素的结构化数据统计开始时间与结束时间均相同, 但是受国内外公共节假日差异等因素的影响, 产生了部分数据的缺失, 本文对于缺失数据采用线性插值的方法进行填补.

2.2 碳交易价格的分解与重构

对 2015 年 1 月 1 日 ~ 2017 年 3 月 10 日的上海碳市场交易价格用 EMD 方法进行分解, 分解得到的 4 个 IMF 和 1 个趋势项, 结果如图 2 所示. 图 2 中: signal 是原始的碳价格序列, imf 1 ~ imf 4 是频率从高到低排列的 4 个本征模函数 (IMF), res 是余下的趋势项序列 $y_3(t)$. 然后, 根据 Fine-to-coarse 方法对分解的子序列进行重构, 重构的结果为 imf 1、imf 2 和 imf 3 共同构成高频序列 $y_1(t)$, imf 4 单独构成低频序列 $y_2(t)$. 图 3 表示对得到的本征模函数重新构造后的高频序列、低频序列、趋势项和原始碳价格序列的走势.

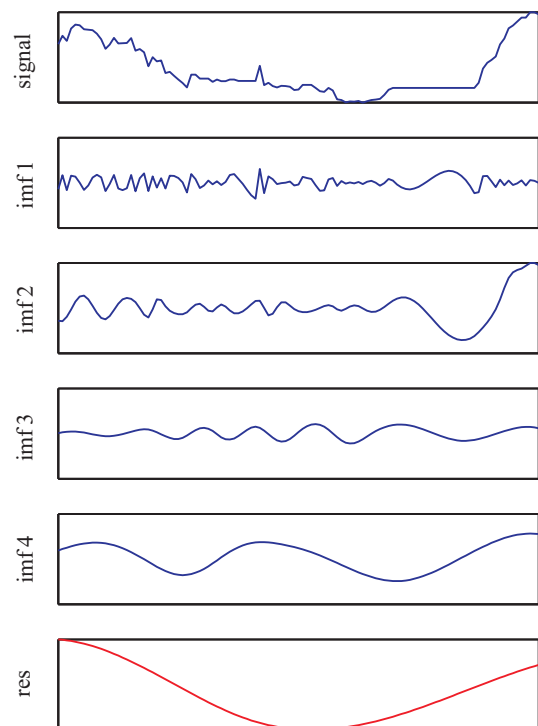


图 2 碳交易价格 EMD 分解结果

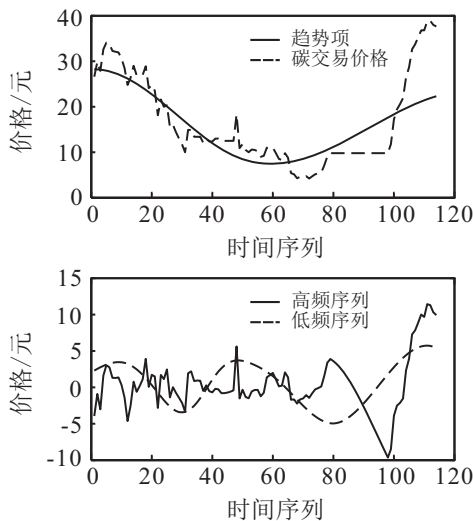


图3 高频率列、低频率列和趋势项的重新构造后的结果

2.3 多尺度组合预测

利用基于等度量映射(Isomap)流行学习的方法对17维的非结构化数据进行降维处理.令 $d = 2$,即保持非结构数据不同时刻的样本点局部欧氏距离不变的情况下,其维度降为2维;再将降维后的非结构化数据和影响因素结构化数据分别进行多尺度分解和重构,形成非结构化数据的高频序列 $z_1(t)$ 、低频序列 $z_2(t)$ 、趋势序列 $z_3(t)$ 和结构化影响因素的高频序列 $s_1(t)$ 、低频序列 $s_2(t)$ 、趋势项序列 $s_3(t)$;然后根据前文的步骤,分别利用不同尺度的有效信息,采用BP神经网络、PLS、ARIMA模型进行多尺度的预测;最后集成3种方法的预测结果即可得到碳价格的预测值.

首先,高频率列的预测运用ARIMA方法,通过1阶差分使高频数据平稳化以确定参数,得到 $p = 2, q = 2$,因此,使用ARIMA(2, 1, 2)模型预测高频率列,得到 $\hat{y}_1(t)$;其次,构造碳价格低频序列 $y_2(t)$ 对 $s_2(t)$ 和 $z_2(t)$ 的PLS回归模型,利用SIMCA-P软件进行PLS回归分析,得到 $\hat{y}_2(t)$;再次,在数据标准化处理的基础上,建立的BP神经网络输入层节点为8,输出层

节点为1,隐含层节点为6,对趋势项序列进行预测,得到 $\hat{y}_3(t)$;最后,将ARIMA预测的高频数据、PLS预测的低频数据、神经网络预测的趋势项数据相加,得到最终组合预测的预测值 $\hat{y}(t)$,如图4所示.由图4可知,基于非结构数据流行学习的多尺度组合预测模型不仅能够有效预测碳交易价格的增长和下降趋势,而且对碳交易价格细节的变动也可以有效拟合.

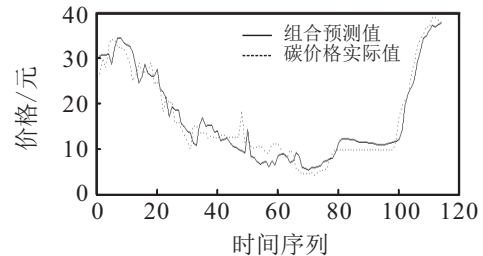


图4 多尺度组合预测模型预测值与实际值比较

2.4 各种预测方法预测效果比较

为了说明本文构建的预测模型的拟合效果,训练集选取2015年1月1日~2016年12月31日的数据,测试集选取2017年1月1日~2017年3月10日的数据.用本文构建的模型与7种预测模型的预测结果进行比较分析.其中:非结构-EMD-ARIMA、非结构-EMD-PLS和非结构-EMD-神经网络这3个模型与本文方法相比,对分解重构后的各频率序列均使用相同的单项预测模型;非结构-组合预测模型则没有对非结构化数据和影响因素结构化数据进行多尺度分解;EMD-组合预测方法没有加入非结构数据;EMD-GARCH模型利用广义自回归条件异方差(GARCH)模型对分解后的碳价格序列进行了预测,为文献[12]所提供的方法;非结构-EMD-最优组合加权与本文方法相比,将高频、低频和趋势项预测值用文献[23]中的基于误差平方和最小的最优加权组合方式进行集成.各方法的预测精度分别使用SSE、MSE、MAE、MAPE、MSPE五种误差评价指标来度量,表1为比较结果.

表1 各模型预测精度的评价指标对比

编号	预测模型	SSE	MSE	MAE	MAPE	MSPE
1	非结构-EMD-ARIMA	2789.97	0.5030	3.8309	0.3066	0.0390
2	非结构-EMD-PLS	1029.26	0.2814	2.3191	0.1755	0.0215
3	非结构-EMD-神经网络	1574.49	0.3481	2.8816	0.2242	0.0289
4	非结构-组合预测	997.83	0.2771	2.3808	0.1917	0.0239
5	EMD-组合预测	834.02	0.2533	2.1290	0.1656	0.0194
6	EMD-GARCH模型 ^[12]	3051.54	0.4889	4.1519	0.3205	0.0390
7	非结构-EMD-最优组合加权	3671.16	0.5315	4.1135	0.3005	0.0367
8	非结构-EMD-组合预测(本文方法)	731.10	0.2372	2.1112	0.1650	0.0193

由表1,通过预测模型1~模型3和模型6与本文方法的比较可以看出,本文所提出预测模型预测结果的SSE、MSE、MAE、MAPE、MSPE这5个误差评价指标值均小于这4种预测模型,说明对各频率序列选择不同的单项预测方法进行预测更合理,优于对各频率序列均使用相同的预测方法.通过预测模型4与本文方法的比较可以看出,多尺度组合-EMD-非结构化数据SSE、MSE、MAE、MAPE、MSPE这5个评价指标值均为最小,表明非结构化数据和影响因素结构化数据经过EMD分解和重构后能避免冗余信息的影响.由模型5与本文方法的比较可以看出,5个评价指标均明显小于模型5,但是就总体预测精度而言,加入非结构化数据的优于不考虑非结构化数据,这进一步验证了基于非结构化数据模型的信息价值.模型7与本文方法进行比较,EMD分解后的高频、低频和趋势项预测值采取组合最优加权集成,由于权重归一化约束的限制,会出现极端权重的情况,不能保证预测的无偏性,可以看出预测精度低于本文方法.综上,仿真实验的结果显示了本文所提出预测方法的优越性,表明本文方法具有较好的拟合效果以及较好的适用性.

3 结论

碳市场交易价格具有非平稳、非线性的特点,针对现有研究没有充分考虑非结构化网络信息的影响,多采用单一模型来预测碳市场的交易价格,预测精度不高等问题,本文在充分利用非结构化数据的基础上,建立了基于非结构数据流行学习的碳价格多尺度组合预测模型.首先,从百度指数提取碳价格相关的非结构化数据,在降维方面使用基于等度量映射的流行学习算法;其次,把降维后的非结构化数据、其他影响因素结构化数据、碳交易价格分别进行EMD分解,得到不同个数的本征模函数(IMF),在此基础上采用Fine-to-coarse技术重新构造以得到高频率序列、低频率序列和趋势项;再次,针对重构后数据序列的不同特点,用ARIMA、PLS回归和神经网络对高频数据、低频数据和趋势项进行预测;最后,将各种预测方法得到的结果集成后得到碳价格序列的最终预测值.仿真实验的结果表明:引入非结构化数据后的多尺度组合预测模型的预测效果明显优于不考虑非结构化数据的效果,验证了非结构化数据对于预测碳价格时间序列具有较高的信息价值;非结构化数据和影响因素结构化数据经过EMD分解和重构处理能避免冗余信息的影响;对各频率序列选择不同的单项预测方法进行预测更合理,优于对各频率序列均使用相同的

预测方法.

碳交易价格的准确预测是提高碳市场风险管理能力的基础,也是制定碳金融市场政策的重要依据.本文提出的多尺度组合预测模型为碳市场交易价格的预测提供了新思路,为人们进行此类研究提供了参考.未来在对碳交易价格进行预测时将会进一步考虑国家政策、国际形势等定性因素的影响.

参考文献(References)

- [1] Chevallier J. Volatility forecasting of carbon prices using factor models[J]. *Economics Bulletin*, 2010, 30(6): 1642-1660.
- [2] Chevallier J. Nonparametric modeling of carbon prices[J]. *Energy Economics*, 2011, 33(6): 1267-1282.
- [3] 张跃军, 魏一鸣. 国际碳期货价格的均值回归: 基于EU ETS的实证分析[J]. *系统工程理论与实践*, 2011, 31(2): 214-220.
(Zhang Y J, Wei Y M. Interpreting the mean reversion of international carbon futures price: Empirical evidence from the EU ETS[J]. *Systems Engineering — Theory & Practice*, 2011, 31(2): 214-220.)
- [4] 张晨, 杨仙子. 基于改进的Grey-Markov对区域碳排放市场价格的预测[J]. *统计与决策*, 2016, 31(9): 92-95.
(Zhang C, Yang X Z. Prediction of regional carbon market price based on improved grey-Markov[J]. *Statistics & Decision*, 2016, 31(9): 92-95.)
- [5] Zhu B, Wei Y. Carbon price forecasting with a novel hybrid ARIMA and least squares support vector machines methodology[J]. *Omega*, 2013, 41(3): 517-524.
- [6] Fan X, Li S, Tian L. Chaotic characteristic identification for carbon price and an multi-layer perceptron network prediction model[J]. *Expert Systems with Applications*, 2015, 42(8): 3945-3952.
- [7] Byun S J, Cho H. Forecasting carbon futures volatility using GARCH models with energy volatilities[J]. *Energy Economics*, 2013, 40(2): 207-221.
- [8] 刘金培, 林盛, 郭涛, 等. 一种非线性时间序列预测模型及对原油价格的预测[J]. *管理科学*, 2011, 24(6): 104-112.
(Liu J P, Lin S, Guo T, et al. Nonlinear time series forecasting model and its application for oil price forecasting[J]. *J of Management Science*, 2011, 24(6): 104-112.)
- [9] 张晨, 杨仙子. 基于多频组合模型的中国区域碳市场价格预测[J]. *系统工程理论与实践*, 2016, 36(12): 3017-3025.
(Zhang C, Yang X Z. Price prediction of China's regional carbon market based on multi frequency combination model[J]. *Systems Engineering — Theory & Practice*,

- 2016, 36(12): 3017-3025.)
- [10] 王书平, 朱艳云. 基于多尺度分析的小麦价格预测研究[J]. 中国管理科学, 2016, 24(5): 85-91.
(Wang S P, Zhu Y Y. Forecasting of wheat price based on multi-scale analysis[J]. Chinese J of Management Science, 2016, 24(5): 85-91.)
- [11] Zhu B, Han D, Wang P, et al. Forecasting carbon price using empirical mode decomposition and evolutionary least squares support vector regression[J]. Applied Energy, 2017, 191(8): 521-530.
- [12] Li W, Lu C. 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.
- [13] 高杨, 李健. 基于EMD-PSO-SVM误差校正模型的国际碳金融市场价格预测[J]. 中国人口、资源与环境, 2014, 24(6): 163-170.
(Gao Y, Li J. Int carbon finance market price prediction based on EMD-PSO-SVM error correction model[J]. China Population, Resources and Environment, 2014, 24(6): 163-170.)
- [14] 张国兴, 张振华, 刘鹏, 等. 我国碳排放增长率的运行机理及预测[J]. 中国管理科学, 2015, 23(12): 86-93.
(Zhang G X, Zhang Z H, Liu P, et al. The running mechanism and prediction of the growth rate of China's carbon emissions[J]. Chinese J of Management Science, 2015, 23(12): 86-93.)
- [15] 赵立祥, 胡灿. 我国碳排放权交易价格影响因素研究——基于结构方程模型的实证分析[J]. 价格理论与实践, 2016(7): 101-104.
(Zhao L X, Hu C. Study on the influence factors of carbon emission trading price in China: An empirical analysis based on structural equation model[J]. Price: Theory & Practice, 2016(7): 101-104.)
- [16] Rausch S, Metcalf G E, Reilly J M. Distributional impacts of carbon pricing: A general equilibrium approach with micro-data for households[J]. Energy Economics, 2011, 33(6): S20-S33.
- [17] 王娜. 基于大数据的碳价预测[J]. 统计研究, 2016, 33(11): 56-62.
(Wang N. Forecasting of carbon price based on big data[J]. Statistical Research, 2016, 33(11): 56-62.)
- [18] Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2009, 457(7232): 1012-1014.
- [19] 董倩, 孙娜娜, 李伟. 基于网络搜索数据的房地产价格预测[J]. 统计研究, 2014, 31(10): 81-88.
(Dong Q, Sun N N, Li W. Real estate price prediction based on web search data[J]. Statistical Research, 2014, 31(10): 81-88.)
- [20] 王炼, 宁一鉴, 贾建民. 基于网络搜索的销量与市场份额预测: 来自中国汽车市场的证据[J]. 管理工程学报, 2015, 29(4): 56-64.
(Wang L, Ning Y J, Jia J M. Predicting sales and market share with online search: evidence from Chinese automobile market[J]. J of Industrial Engineering and Management, 2015, 29(4): 56-64.)
- [21] 陈植元, 米雁翔, 厉洋军, 等. 基于百度指数的投资者关注度与股票市场表现的实证分析[J]. 统计与决策, 2016(23): 155-157.
(Chen Z Y, Mi Y X, Li Y J, et al. Investor concern based on the Baidu index an empirical analysis of the performance of the stock market[J]. Statistics & Decision, 2016(23): 155-157.)
- [22] 颜七笙, 王士同, 张延飞, 等. 基于经验模式分解和极限学习机的铀资源价格预测方法[J]. 控制与决策, 2014, 29(7): 1187-1192.
(Yan Q S, Wang S T, Zhang Y F, et al. Uranium resource price prediction based on empirical mode decomposition and extreme learning machine [J]. Control and Decision, 2014, 29(7): 1187-1192.)
- [23] 陈华友. 组合预测方法有效性理论及其应用[M]. 北京: 科学出版社, 2008: 60-64.
(Chen H Y. The theory of the effectiveness of combined forecasting method and its application[M]. Beijing: Science Press, 2008: 60-64.)
- [24] 陈晓红, 王陟昀. 碳排放权交易价格影响因素实证研究——以欧盟排放交易体系(EUETS)为例[J]. 系统工程, 2012, 30(2): 53-60.
(Chen X H, Wang Z Y. Empirical research on price impact factor of carbon emission exchange: Evidence from EU ETS[J]. Systems Engineering, 2012, 30(2): 53-60.)
- [25] Tenenbaum J B, De S V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2322.

(责任编辑: 李君玲)