Forecasting Price Trend of Bulk Commodities Leveraging Cross-domain Open Data Fusion

BINBIN ZHOU and SHA ZHAO, Zhejiang University, China
LONGBIAO CHEN, Xiamen University, China
SHIJIAN LI, ZHAOHUI WU, and GANG PAN, Zhejiang University, China

Forecasting price trend of bulk commodities is important in international trade, not only for markets participants to schedule production and marketing plans but also for government administrators to adjust policies. Previous studies cannot support accurate fine-grained short-term prediction, since they mainly focus on coarse-grained long-term prediction using historical data. Recently, cross-domain open data provides possibilities to conduct fine-grained price forecasting, since they can be leveraged to extract various direct and indirect factors of the price. In this article, we predict the price trend over upcoming days, by leveraging cross-domain open data fusion. More specifically, we formulate the price trend into three classes (rise, slight-change, and fall), and then we predict the specific class in which the price trend of the future day lies. We take three factors into consideration: (1) supply factor considering sources providing bulk commodities, (2) demand factor focusing on vessel transportation with reflection of short time needs, and (3) expectation factor encompassing indirect features (e.g., air quality) with latent influences. A hybrid classification framework is proposed for the price trend forecasting. Evaluation conducted on nine real-world cross-domain open datasets shows that our framework can forecast the price trend accurately, outperforming multiple state-of-the-art baselines.

CCS Concepts: • **Information systems** → **Data mining**; • **Applied computing** → **Marketing**; **Forecasting**; Economics;

Additional Key Words and Phrases: Price trend, cross-domain data, data fusion, multi-class prediction, bulk commodity

ACM Reference format:
Binbin Zhou, Sha Zhao, Longbiao Chen, Shijian Li, Zhaohui Wu, and Gang Pan. 2020. Forecasting Price Trend of Bulk Commodities Leveraging Cross-domain Open Data Fusion. *ACM Trans. Intell. Syst. Technol.* 11, 1, Article 1 (January 2020), 26 pages.
https://doi.org/10.1145/3354287

This work is supported by Ten Thousand Talent Program of Zhejiang Province (Grant No. 2018R52039) and National Natural Science Foundation of China (Grant No. 61772460).

Authors' addresses: B. Zhou, S. Zhao, S. Li (corresponding author), Z. Wu, and G. Pan, College of Computer Science and Technology, Yuquan Campus, Zhejiang University, 38 Zheda Road, Hangzhou, 310000, China; emails: {bbzhou, szhao, shijianli, wzh, gpan}@zju.edu.cn; L. Chen, Room 601, Administration Building, Haiyunyuan, Xiamen University, Xiamen, 361005, China; email: longbiaochen@xmu.edu.cn.

2157-6904/2020/01-ART1 $15.00
https://doi.org/10.1145/3354287

1 INTRODUCTION

Bulk commodities play a fundamental part in a country's economic development, which refer to raw materials transported in large quantities for production and trade, such as iron ore, coal, and grain. It is important to forecast prices of bulk commodities accurately, not only for purchasers and suppliers but also for traders and even government administrators. Purchasers can adjust manufacturing plans and purchase bulk commodities in lower prices to maximize profits. Suppliers can schedule operation and marketing plans to control the supply-demand balance. Traders can purchase these commodities from large-scale overseas suppliers and sell to small-scale domestic companies with urgent demand, taking advantage of a long seaborne delivery time. For government administrators, normally, they would prefer long-term bulk commodities price forecast to monitor and adjust their resource policies. They could also consider short-term forecast to micro-adjust the local resource policies and provide financial assistant for related companies.

However, it is difficult to accurately forecast prices of bulk commodities. Here, we take one typical and important bulk commodity, iron ore, as an example, to explain the difficulty. Iron ore is a dry bulk commodity, with the largest trade volume per year [30]. It is commonly used for steel production as key ingredients, and it is irreplaceable in steel manufacturing due to its homogeneous characteristic nowadays [22]. Market participants trade iron ore in a spot price based on a well-acknowledged price index Platts Iron Ore Index (IODEX) [33]. The IODEX provides meaningful and transparent representation of physical iron ore market. The confirmation of IODEX relies on market participants' heuristic bids and offers, resulting in a difficulty in estimation and prediction. Specifically, Platts workers collect various data about bids, offers, expressions of interest to trade, and confirmed trades and then publish the latest data on the Market on Close (MOC) as the corresponding IODEX price. All these bids, offers, and expressions of interest would change dynamically and heuristically, and they are also mixed with participants' psychological expectations. Hence, the price is affected by various factors. Besides, the fact that it is not easy to obtain some critical relevant data for the commercial secrets and limited data access, makes it difficult to forecast the price accurately. In this article, for simplification purposes, we model the price forecast problem as a classification task, i.e., forecasting the price trend of bulk commodities.

Existing literature on prices of bulk commodities mainly focuses on long-term price trend forecast, paying less attention to the short-term price trend forecast. Due to the coarse-grained data collected and processed, it is hard to conduct fine-grained short-term analysis and forecast. Moreover, previous studies on bulk commodities-related analysis mainly rely on economic-specific statistical data [21, 26, 40]. For bulk commodities, using specific-domain data is insufficient to discover some latent factors of prices, since the price changing is influenced by various kinds of factors from different domains. For example, air quality of steel-intensive cities affected by steel production, reflect latent demand of iron ore and coal, thus bring influences on these price trends.

Although the growing availability of open datasets from various domains provides us an opportunity to make short-term price trend forecast for bulk commodities, there still remain two challenges to forecast the price trend accurately. First, there is limited access of data directly related to the price trend, due to trade secrets and lack of unified data collection platforms. Second, combined with the high variety of open data, it is hard to confirm and extract directly representative factors related to the price trend.

In this article, to address these aforementioned challenges, we attempt to leverage cross-domain open data to forecast the price trend of bulk commodities over upcoming days. Here, the price trend is defined as three classes, i.e., rise, slight-change and fall. We first analyze three influential factors, i.e., supply factor, demand factor, and expectation factor. Then, we identify and select relevant features from cross-domain open datasets according to these three factors. After that, we feed selected features into a hybrid prediction model consisting of multiple classification models, to achieve a sequential three-class forecast of following T days. Finally, we conduct performance evaluation utilizing real-world datasets. Experimental results show that our framework can accurately predict the price trend in three classes following T business days, outperforming all state-of-the-art baselines. The main contributions of our article include:

(1) Our work is a promising step towards short-term price trend forecast of bulk commodities leveraging cross-domain open data fusion. We employ open data both from physical domains (e.g., vessel trajectory and air quality), and economic-specific domains (e.g., seaborne cost and stock price), to conduct a cross-domain study. The results demonstrate that it is a novel and successful attempt for cross-domain studies utilizing open data.

(2) We propose a three-layer framework for a sequentially continuous T days' price trend prediction: cross-domain open data layer, feature selection layer, and price trend prediction layer. In the cross-domain open data layer, we collect all possible and accessible open data from various domains. In the feature selection layer, we first identify three related key factors from various cross-domain open data sources: supply factor from port statistical data, demand factor mainly from port statistical data and trajectory data, and expectation factor from urban data and economic data. We then select most relevant features from all combined factors for each prediction time interval. In the price trend prediction layer, we propose a hybrid model consisting of multiple models to achieve a sequential three-class prediction.

2 RELATED WORK

In this section, we review the relevant previous work from two viewpoints. We investigate the studies forecasting economic indicators, including prices of bulk commodities, and introduce the existing literature using cross-domain open data.

2.1 Economic Indicators Prediction

There are a number of studies on forecasting various kinds of economic indicators in different scenarios, such as iron ore price, stock index in different countries or regions, currency in different countries and electricity prices. For example, References [8, 49] leveraged historical hourly electricity prices to predict the next-day price using different methods, respectively. Contreras et al. [8] employed ARIMA to model moving trends of electricity prices and then predict the future price. Implementations on mainland Spain and Californian markets verified the effectiveness of this study. Yadav et al. [49] proposed a hybrid method for accurate electricity price prediction by employing fuzzy systems in the Standard PSO method and applied the method to the electricity markets of Spanish to validate its superiority over many baseline methods. Wang [42] studied the stock price prediction problem with nonlinear neural networks. The author used historical trade prices to predict future stock price, and experiments on Taiwan Stock Index demonstrated the proposed method can obtain accurate prediction. These aforementioned studies on economic indicators prediction usually conduct short-term predictions due to their large amount of fine-grained historical data, e.g., hourly and daily data. However, it is difficult to conduct short-term fine-grained prediction of bulk commodities with only historical specific-domain data due to the coarse-grained bulk commodities data.

Compared with other products, such as electricity prices and stock prices, the short-term price prediction of bulk commodities has its intrinsic characteristics. First, the trade of bulk commodities usually involves international export and import, encompassing complex factors affecting their prices changing. For example, the different regulatory frameworks in different trade-related countries would increase the trade complexity of bulk commodities. Second, the data is not dense with smaller samples. Bulk commodities are usually traded in large quantities with lower frequency, while other products (e.g., stocks) are usually traded with higher frequency. In this way, relevant data of bulk commodities are sparser. Thus, the short-term price prediction for bulk commodities is quite challenging. In the past decades, previous literature on bulk commodities-related studies usually involved various datasets. For instance, Pustov et al. [34] conducted long-term (more than 5 years) iron ore price prediction study using historical iron ore price data and various factors, e.g., operating costs, investment return, and demand growth. Zhang et al. [50] investigated the monthly soybean price prediction using various datasets, including historical monthly soybean prices, the output of domestic and global soybean, input volume of soybean, domestic demand of soybean and so on. They employed a quantile regression-radial basis function neural network to solve this problem, and experiments results verified the advantages of the proposed method, which can accurately predict the soybean prices. Although these bulk commodities related studies have incorporated some relevant data to improve the price prediction performance, they are unable to obtain accurate and short-term price prediction, due to the limitation of coarse-grained historical data. In this article, we attempt to conduct short-term forecast on price trend of bulk commodities leveraging datasets from various domains.

2.2 Cross-domain Open Data Applications

Recently, there is an increasing trend for applying cross-domain open big data to address pervasive challenges, e.g., intelligent transportation planning and optimization [2, 3, 5, 7, 17, 31, 54], smartphone users profiling and understanding [15, 51–53], and urban environment monitoring [37, 38, 55]. For instance, Zheng et al. [55] predicted fine-grained air quality readings of monitoring stations during the next two days in a real-time manner, by using various datasets from different domains, e.g., historical air quality data of all cities involved, meteorology data, and weather prediction data. Chen et al. leveraged various cross-domain urban open data to conduct bike sharing-related studies. They proposed a semi-supervised framework to learn and rank bike sharing stations using POIs, Check-in, and demographics data in Reference [5], and predicted dynamic cluster-based over-demand of bike sharing stations using weather condition, air temperature, social events and traffic events data in Reference [7]. Xiong et al. [16] ranked residential real estates with the utilization of multiple cross-domain datasets, such as online user reviews (e.g., overall satisfaction score and service quality score), and offline moving behaviors (taxi trajectory, smart card transactions, and check-in data).

However, in the past decades, there were few studies on economic indicators prediction using cross-domain open data. Fortunately, ubiquitous open data from various domains provide us an unprecedented opportunity to conduct economic indicators prediction studies leveraging cross-domain open data fusion. It is not trivial to predict economic indicators using cross-domain data. Intuitively, we usually apply the supply and demand-related data as factors to influence the economic indicator. Nevertheless, it is hard to obtain supply and demand-related data directly. Leveraging relevant cross-domain data to help represent latent factors becomes promising and practical. For example, Chen et al. [4, 6] conducted studies on container port performance measurement and comparison. They used ship GPS trajectory record, to identify container handling events. They then used cross-domain data, e.g., ship GPS trajectory and open port facilities data, to estimate containers number of multiple ports. Therefore, in this article, we explore the possibility to incorporate various cross-domain open data to study the short-term price trend forecasting problem.

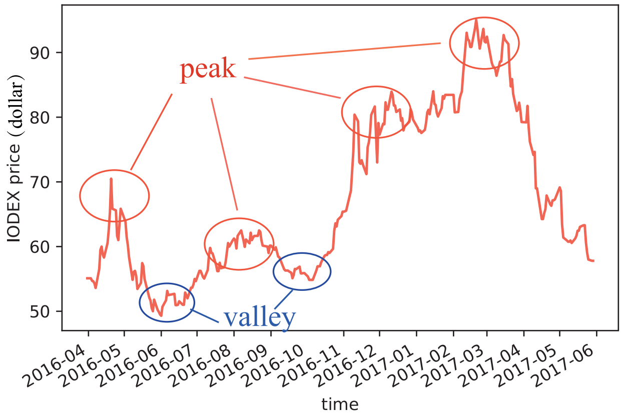
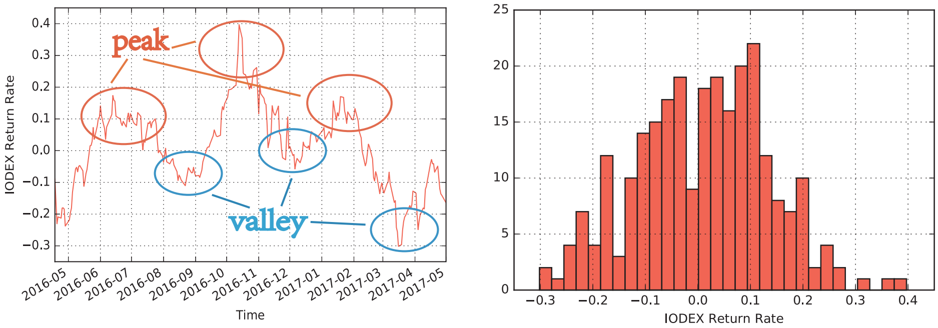


Fig. 1. Example: Iron ore price trend from April 2016 to May 2017.



(a) change of return rate

(b) distribution of return rate

Fig. 2. IODEX return rate when $\Delta T = 20$.

3 PROBLEM DESCRIPTION AND FRAMEWORK OVERVIEW

3.1 Problem Description

The objective of this work is to forecast the price trend of bulk commodities over multiple future days, by utilizing cross-domain open datasets. Sensing price change in advance can help them increase profits and schedule plans. Therefore, we formulate the price trend into three classes: rise, slight-change, and fall. This problem, thus, can be represented as a three-class prediction problem. Three labels are defined corresponding to the three classes, -1, 0, and 1, where -1 denotes a fall of the iron ore price, 0 denotes a slight-change of the price, and 1 denotes a rise of the price.

We take iron ore, one typical bulk commodity, as a representative to present the problem formulation. We first collect historical iron ore price index (IODEX price) data to observe its changing trend. From Figure 1, we observe that the price curve has an obvious changing from April 2016 to May 2017, with four peaks and two valleys. With these sharp changes, it is not easy to predict the specific class of future price trend lies in. The most obvious reason may be that the price moves in violent fluctuation, even in a small-range peak duration or valley duration. Furthermore, forecasting future price trend using historical price data, only interprets its economic significance in a coarse-grained way, without paying attention on surrounding related data from other domains. Therefore, we use various open data from different domains for this three-class price trend problem, to forecast whether the price would rise, be in slight-change or fall.

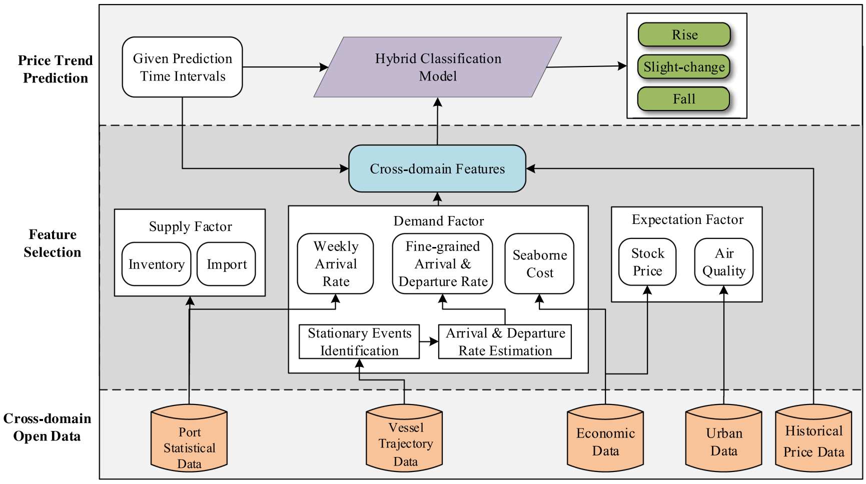


Fig. 3. Overview of the framework.

We divide the price trend into three classes based on its return rate, which can be computed in Equation (1). Here, p_s and p_e are defined as iron ore price values of start time s and end time e, respectively. $\psi(e,s)$ is defined as the return rate from s to e. Computation results are presented in Figure 2(a). We observe that values of the curve are from −0.3 to 0.4, with some inflection points and some slight-change points. We further discretize them, as shown in Figure 2(b). We then define three classes using a threshold δ, i.e., the return rate is (1) higher than δ (class *rise*), (2) within the range $[-\delta, \delta]$ (class *slight-change*), and (3) lower than $-\delta$ (class *fall*). We attempt to comprehensively interpret the price trend from different aspects:

$$\psi(e,s) = (p_e - p_s)/p_s. \tag{1}$$

3.2 Framework Overview

Based on the aforementioned analysis, we explore the price trend forecast over multiple future days utilizing various open data from different domains. The big variant cross-domain data brings both relevant and irrelevant features, the latter would play negative roles in the prediction process. In this way, feature selection is necessary and should be taken into account. Therefore, we propose a three-layer framework for the price trend prediction: (1) cross-domain open data layer, (2) feature selection layer, (3) price trend prediction layer, and we show it in Figure 3. This framework is flexible and can be generalized to predict the price trend of other bulk commodities. We employ one representative of bulk commodities, iron ore, to describe the framework in the following.

(1) **Cross-domain Open Data**. We first retrieve all possible relevant data according to the analysis of previous studies and prior knowledge. Then, we collect all possible and accessible open data from various domains, such as historical price data, inventory quantity and import quantity data, trajectory data of vessels transporting seaborne iron ore, urban air quality data of cities in China, and stock price of relevant enterprises. Data from various domains can help interpret the price trend from different aspects, and help improve the prediction accuracy.

(2) **Feature selection**. We identify three influential factors, i.e., supply, demand and expectation. To effectively quantify the influence of each dataset for the prediction, we then categorize all collected open datasets into the three non-overlapping factors. The supply

and demand factors focus on the formulation of the price. Specifically, the supply factor mainly relies on datasets, indicating sources providing iron ore, e.g., iron ore inventory in coastal ports and iron ore import from the overseas regions. For the demand factor, we considers iron ore demand and consumption in short time in the way of vessel seaborne transportation, e.g., arrival vessels in coastal ports, arrival/departure rate in specific relevant ports and seaborne cost. The expectation factor mines the further latent features for the price. For example, air quality can be affected by the usage of iron ore and reflect the future demand of iron ore. Combining the three factors' data with historical iron ore price data, we select different relevant features corresponding to different prediction time intervals.

(3) **Price Trend Prediction**. We feed all selected features to generate different datasets according to different future time intervals. With the input data, we apply a hybrid classification model consisting of multiple prediction methods, to forecast a specific class of the price over following days, rise, slight-change or fall. Through extensive experimental results, we compare the performance of state-of-the-art prediction methods, to explore suitable models for the price trend prediction over upcoming days. Furthermore, some prediction time intervals with accurate and acceptable prediction results also can be recognized. The results would be beneficial for market participants to help them schedule plans and adjust policies.

4 ANALYSIS OF PRICE TREND FACTORS

To understand influential factors of the price trend, researchers have conducted a series of studies [19, 34]. Based on the prior knowledge, we identify the following factors in determining the price trend, i.e., supply, demand, and expectation. The more bulk commodities supplied with stable demand would bring fierce competitions among traders and sellers, leading to decreasing prices and corresponding profits. In contrast, the more bulk commodities demand with stable supply also would change the equilibrium of prices. Another type of competitions among traders and sellers would occur, resulting in increasing prices and profits. However, the determination of prices are not easy. First, the supply and demand are dynamically changing with different time. Second, there are also some marginal and extra potential influential factors, such as market participants' confidence. We name these extra influential factors as expectation factors. Based on the above analysis, we select a set of open datasets related to the three factors, and then conduct correlation analysis of each group factors with iron ore as an example of bulk commodities.

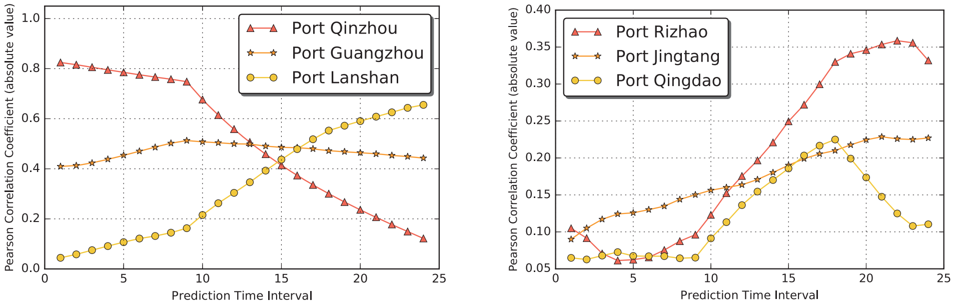
4.1 Supply Factor

There are three sources continuously providing iron ore for Chinese purchasers, including iron ore inventory in ports owned by trading companies, iron ore import from overseas mining companies, and domestic iron ore mined and sold by domestic companies. Among them, the domestic iron ore has lower competitiveness due to its lower grade in iron ore and consequently higher processing costs. Steel mills prefer iron ore from Australia and Brazil, with high grade in iron ore and low processing costs. Consequently, iron ore is usually traded in the way of iron ore inventory in coastal ports or direct overseas import. Therefore, the supply factor includes two groups of data: the inventory quantity data and import quantity data of the past d days in p coastal ports. We calculate relevant data from April 18, 2016 to April 20, 2017. We then analyze these data as follows.

4.1.1 Iron Ore Inventory. We analyze the correlation between inventory quantity and iron ore price trend to verify their relevance. Specifically, for each port we first retrieve the inventory quantity of all available ports collected by each Friday. We then compute the Pearson correlation

Table 1. Correlation Coefficients of Top-10 Ports Most Relevant to IODEX Trend w.r.t. Inventory

	$\Delta T = 1$	$\Delta T = 5$	$\Delta T = 10$	$\Delta T = 20$
1	Qinzhou (−0.8247)	Luojing (0.8335)	Luojing (0.7859)	Lanshan (−0.5906)
2	Changzhou (0.8245)	Changzhou (0.8141)	Changzhou (0.7451)	Nantong (0.4713)
3	Zhanjiang (0.8240)	Zhanjiang (0.8094)	Zhangjiagang (0.7313)	Guangzhou (0.4642)
4	Luojing (0.8105)	Qinzhou (−0.7855)	Fangchenggang (0.7209)	Zhajiagang (0.4621)
5	Rizhao (0.7485)	Zhangjiagang (0.7462)	Zhanjiang (0.7116)	Fanchenggang (0.45)
6	Lianyungang (0.7385)	Fangchenggang (0.7231)	Qinzhou (−0.6768)	Beilun (0.4356)
7	Jiangyin (0.7150)	Lianyungang (0.7132)	Lianyungang (0.6045)	Luojing (0.4272)
8	Zhangjiagang (0.7138)	Rizhao (0.7089)	Longkou (0.5746)	Changzhou (0.3674)
9	Fangchenggang (0.6880)	Jiangyin (0.6811)	Nantong (0.5718)	Zhanjiang (0.2587)
10	Tianjin (0.6521)	Longkou (0.6418)	Rizhao (0.5705)	Qingdao (0.2402)



(a) Inventory (b) Import

Fig. 4. Changing trend of correlation coefficient of supply factors.

coefficient to measure their relationship, respectively. Table 1 shows the top-10 ports most relevant to the IODEX trend w.r.t. iron ore inventory. Here, ΔT is defined as the prediction time interval. From the table, we can find that port Qinzhou has a negative correlation coefficient while most of other ports have positive correlations. Intuitively, the price trend would vary with the balance between supply and demand. When the demand is greater than the supply, the price has high possibility to rise. China has increased investment on infrastructure construction by 17.4% in 2016 and 19% in 2017, respectively [28]. With this high demand of iron ore, the price keeps rising while the inventory quantity in each ports keep rising as well. Hence, for the ports having positive correlations with the price trend, one possible reason is that the demand of iron ore is higher than the supply continuously during this period of time. Besides, for port Qinzhou having a negative correlation with the price trend, we find that the port congestion issue merely happens in the small-scale port Qinzhou, which usually has inventory quantity as 10–30 million tons. One possible reason may be that the local steel mills consume iron ore quickly, reflecting the urgent demand of iron ore directly. So when the inventory quantity in Qinzhou decreases, the iron ore price rises. Besides, we need to notice that, ports having high Pearson correlation coefficients with IODEX trend in iron ore inventory quantity, may not keep advantages with ΔT increasing.

Furthermore, we explore the impact of inventory on IODEX trend in different prediction time intervals successively, and present the correlation coefficient changing trend of three selected ports in Figure 4(a). From the figure, it can be seen that different ports have different changing trends. We observe that the inventory quantities in some ports have gradually decreasing correlations with

Table 2. Correlation Coefficients of Top-6 Ports Most Relevant to IODEX Trend w.r.t. Import

	$\Delta T = 1$	$\Delta T = 5$	$\Delta T = 10$	$\Delta T = 20$
1	Rizhao (0.1048)	Jingtang (0.1258)	Jingtang (0.1563)	Rizhao (0.3460)
2	Jingtang (0.0899)	Caofeidian (0.0934)	Caofeidian (0.1306)	Caofeidian (0.2384)
3	Caofeidian (0.0857)	Qingdao (0.0674)	Rizhao (0.1230)	Jingtang (0.2245)
4	Qingdao (0.0648)	Rizhao (−0.0623)	Qingdao (0.0912)	Qingdao (0.1736)
5	Tianjin (−0.0459)	Tianjin (−0.0269)	Lianyungang (0.0384)	Tianjin (0.1560)
6	Lianyungang (0.0053)	Lianyungang (0.0259)	Tianjin (0.0006)	Lianyungang (0.1209)

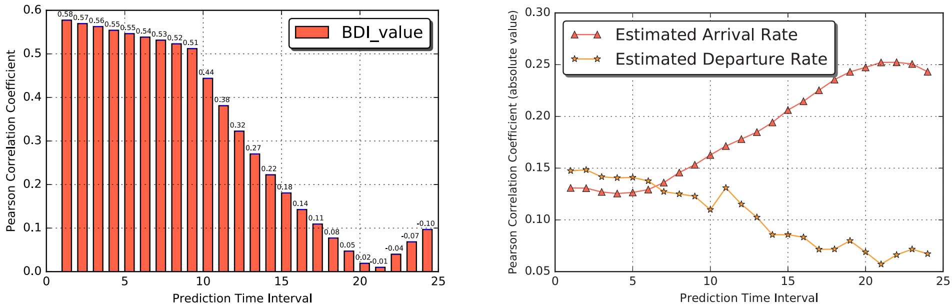
the price trend when ΔT grows, such as Port Qinzhou. These ports are usually small scale ports. The iron ore inventory can be consumed quickly by local steel mills. In this way, the inventory quantities in these ports have diminishing correlations with the iron ore price. Moreover, we also observe that the inventory quantities in some ports have increasing correlations with the price trend when ΔT grows, such as Port Lanshan. These ports are usually large scale ports, usually having inventory quantities greater than 300 million tons. These iron ores are usually transported from large overseas iron ore mining companies. So the changing trend of correlation in these ports may reflect that the latent changing national demand of iron ore. Therefore, the inventory quantities in these ports have growing correlations with the iron ore price. Besides, we find that the inventory quantities in some ports keep relative stable correlations with the price trend when ΔT grows, such as Port Guangzhou. These ports usually located at significant coastal areas. Some of these iron ore inventories are usually transported to some small scale inland ports for local steels mills and trading companies. Hence, using only the inventories in these ports is not easy to further clarify their correlations with the price directly when ΔT grows. Therefore, we introduce more data sources for further analysis, e.g., iron ore import in ports.

4.1.2 Iron Ore Import. We also characterize the iron ore import part by the import quantity collected from available ports, and analyze their correlations. Due to the limited data resources, we calculate six major relevant ports data, and present the analysis results in Table 2. Note that import quantities in all the ports have very weak relationship with IODEX trend. The correlation coefficient of import quantity in each port with the price trend varies when ΔT changes. We analyze the impact of import on IODEX trend in different ΔTs as well, and present it in Figure 4(b). From the figure, we can observe that the trends of correlation coefficient between import quantity in different ports and iron ore price vary with ΔT grows, similar to the iron ore inventory.

4.2 Demand Factor

Based on the influence caused by iron ore supply, the price is also susceptible by iron ore demand. But due to the trade secrets, it is hard to obtain the iron ore demand plans openly from potential purchasers, e.g., steel mills and iron ore traders. Meanwhile, the iron ore desired by major purchasers (e.g., China) will be transported from the major overseas suppliers (e.g., three predominant iron ore producers). In this way, we consider to analyze the demand factor from the perspective of vessel transportation. Specifically, we take two critical factors into account, seaborne cost and vessels mobility, to reflect the vessel transportation situation.

4.2.1 Seaborne Cost. For seaborne cost, a high demand should be reflected in a high shipping cost. More specifically, the demand of iron ore is driven by the endogenous need of the purchasers, e.g., steel mills and trading companies. And then, high demand will result in a more competing shipping market, and thus raise the seaborne cost. Therefore, we select seaborne cost to analyze



(a) BDI

(b) Arrival & Departure Rate

Fig. 5. Changing trend of correlation coefficient of demand factors.

Table 3. Correlation Coefficients of Top-6 Ports Most Relevant to IODEX Trend w.r.t. Vessel Arrival Rate

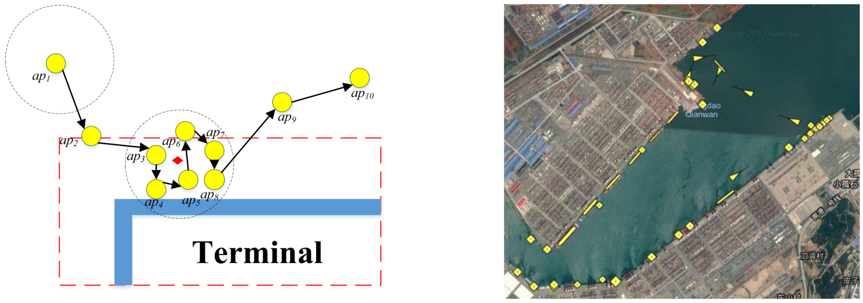
	$\Delta T = 1$	$\Delta T = 5$	$\Delta T = 10$	$\Delta T = 20$
1	Caofeidian (0.1840)	Caofeidian (0.1650)	Caofeidian (0.1784)	Rizhao (0.2886)
2	Rzhao (0.1455)	Jingtang (0.1520)	Jingtang (0.1674)	Caofeidian (0.2323)
3	Qingdao (0.1298)	Qingdao (0.1289)	Rizhao (0.1580)	Jingtang (0.2061)
4	Jingtang (0.1162)	Rizhao (0.1010)	Qingdao (0.1374)	Qingdao (0.2053)
5	Lianyungang (0.0378)	Tianjin (−0.0253)	Lianyungang (0.0654)	Lianyungang (0.1633)
6	Tianjin (−0.0278)	Lianyungang (0.0123)	Tianjin (−0.0133)	Tianjin (0.1311)

the demand factor of iron ore. Here, we characterize the seaborne cost by employing the Baltic Dry Index (BDI)[1], a daily economic indicator to assess the seaborne cost of bulk commodities. We then analyze its correlation with IODEX trend in different ΔTs successively. From Figure 5(a), we observe that BDI has decreasing correlation with iron ore price when ΔT increases. It has high correlation with iron ore price, and achieves 0.58 when $\Delta T = 1$. When $\Delta T = 21$, the correlation reaches the weakest point as minus 0.01, and then starts growing the correlation in a negative manner. Therefore, we need to introduce more datasets to find relevant features when ΔT grows.

4.2.2 *Vessel Arrival and Departure Rate.*

For vessels mobility, the arrival rate of vessels transporting iron ore should be considered, due to its indication of vigorous demand of iron ore in short time. Currently, the accessible and easy-access data only provide weekly arrival rate of only 6 coastal ports in China. We analyze the impact of weekly arrival rate of vessels on IODEX trend in different ΔTs, and demonstrate it in Table 3. From the table, we can find that vessel arrival rate in these ports have weak relationship with IODEX trend. And the correlation coefficient of vessel arrival rate in each port with the price trend varies with ΔT. With ΔT increases, vessel arrival rates in some ports have growing correlation with IODEX trend, such as Port Rizhao. High vessel arrival rate may reflect high vessel mobility. The competing shipping market would be influenced, reflecting the latent demand of iron ore as well.

We notice that the transportation duration from overseas mining regions to China would reach 15 to 30 days. It would be beneficial for our study to conduct a fine-grained estimation of arrival rate to complement the weekly arrival rate, and also estimation of departure rate of overseas ports with major iron ore export. For this issue, we notice that the trajectory of these specific types of vessels can complement and refine the weekly arrival rate. In addition, the trajectory data can also extract the number of departure vessels in the departure ports located in iron ore production



(a) vessels in a terminal (b) vessels in Port Qingdao

Fig. 6. Illustration of vessels in ports and terminals.

regions. In this way, we decide to collect all vessels' data who participate in iron ore seaborne transportation.

Due to the constraint of iron ore transportation, there are specific types of vessels for the iron ore seaborne transportation, including *very large ore carriers*, *capsize bulk carriers*, *paramax bulk carriers*, *handymax bulk carriers*, and *handysize bulk carriers*. The type of *Very large ore carriers* often falls into the *capsize bulk carriers* group, and here we also accept this popular and general operation. According to the vessel name, we can confirm and collect their trajectory record data. On this process, we may encounter some vessels with a same vessel name. When meet these situation, we use the dead-weight ton (DWT) to narrow down and confirm the unique desired vessel. In this way, we collect the trajectory record data of all involved vessels.

Arrival and Departure Rate Refinement. We propose a two-phase model to estimate arrival and departure rate of relevant ports leveraging vessels' trajectory record data. The first phase is stationary event identification, in which we take various features from trajectory data into account, i.e., location, traveling speed and heading direction, to detect the stationary event. The second phase is arrival and departure rate estimation, in which we confirm whether any vessel with stable location is located in any relevant ports, including ports in Australia, Brazil, and China.

Phase I —Stationary Event Detection. The readings of a vessel's position is not solid due to the AIS errors, even when the vessel stays stationary without any moving [4]. Thus, it is necessary to extract stationary events from the large amount of AIS trajectory data. A vessel AIS trajectory tra is a sequence of AIS points ap that the movement of an object recoded in the format of latitude lat, longitude lng, time stamp t, heading hd, speed sp with the identification of the object vid. Usually, $ap.vid$ is confirmed by the maritime mobile service identify (mmsi) of the vessel. In this way, a vessel's trajectory can be represented as $tra = ap_1 - > ap_2 - > \cdots - > ap_n$.

We employ an adaptive sliding window-based approach to detect all possible stationary events [4, 56], with both distance and time difference constraints. As shown in Figure 6(a), for a tra, we start by checking ap_m-> $\cdots - > ap_{m+k}$, whether both $ap_{m+i}.sp < \varepsilon$ and $ap_{m+i+1}.sp < \varepsilon$. If $ap_{m+i}.sp >> 0$ or $ap_{m+i+1}.sp >> 0$, then it indicates the vessel is moving, rather than berthing or even staying stationary. If $ap_{m+i}.sp < \varepsilon$ and $ap_{m+i+1}.sp < \varepsilon$, then we continue to check whether $Distance(ap_{m+i}, ap_{m+i+1}) < \Delta d$ and $|ap_{m+i}.t - ap_{m+i+1}.t| > \Delta t$. If so, then it indicates that the vessel moves in a small range during a long time, which is most possible for berthing or staying stationary. Thus, we expand the window size to cover one more ap until a new coming ap_{m+j} has a larger distance to the first point in the window than Δd, no matter the time difference between the two ap. In the experiment, we set Δd = 1 meter and Δt = 5 minutes for stationary event detection.

Phase II—Arrival and Departure Rate Estimation. For all detected stationary events, we represent the stationary event $event$ as $event = (ap_s, sp_e, t_s, t_e, lat_s, lat_e, lng_s, lng_e)$. Here, s denotes the start time, e denotes the end time, lat denotes the latitude, and lng denotes the longitude. Note that there are some events our problem does not cover. For example, some temporary stays happen at sea, and some berthing events occur in other ports rather than our desired ports, like vessels from Brazil to China stopping in Indonesia and Philippines. Here, we need to confirm original and destination ports of iron ore transportation. We confirm the original ports as port Hedland, port Dampier, port Walcott, port Tubarao, port Ponta Da Madeira, and port Itaguai, since the typical iron ore sellers are located in Western Australia and Brazil. For destination ports, we select the coastal ports in China as the destinations, such as port Qingdao. And then, we check their position with a rectangle area $area$ to cover the port, represented as $area = (lat_1, lat_2, lng_1, lng_2)$. With the several predefined rectangle areas, we identify whether the stationary $event$ occurs in these areas, if $\prod_{i=1}^{2}(lat_i - lat_j) < 0$ and $\prod_{i=1}^{2}(lng_i - lng_j) < 0, j = s, e$.

As shown in Figure 6(b), we present a snapshot of port Qingdao in China by satellites with multiple vessels travelling in the sea and staying in the port. We only consider arrival events happened in destination ports and departure events occurred in original ports. After the identification of the arrival and departure events of original and destination ports, we are able to estimate daily arrival rate and departure rate by summing up the arrival or departure events.

With these estimated arrival rate and departure rate, we analyze the changing trend of their correlation coefficient with IODEX trend in different ΔTs, as shown in Figure 5(b). From the figure, it can be seen that the estimated arrival rate and departure rate have different changing trends. We observe that the estimated arrival rate has increasing correlations with the price trend when ΔT grows. The changing trend of correlation coefficient of estimated arrival rate has verified our aforementioned analysis on weekly arrival rate, and the necessity of this fine-grained estimation. Moreover, we observe that the estimated daily departure rate has decreasing correlations with the price trend when ΔT grows. After several shipping days, these vessels from overseas ports with major iron ore export reach destination ports, bringing a large number of bulk commodities (e.g., iron ore). At that time, some demand of iron ore would be satisfied. Hence, the changing trend of correlation coefficient of estimated departure rate decreases with ΔT.

4.3 Expectation Factor

Besides these direct effects, there are also some indirect influences. We identify two critical features based on our prior knowledge, air quality and stock price. When the air quality of some cities gets worse, with high concentration SO2, NO2, and CO, we should pay attention to these pollutions. The emissions of these high concentration pollutants are usually from the steel and iron manufacturing processes [41, 48]. These high concentration pollutants may indicate the demand of iron ore not only in this period but also in the short future. In addition, the stock price of market entities involved in this industry, including steel mills, trading companies, port companies, and seaborne-related companies, also represents the prosperity of this industry and market investors' confidence. Hence, we retrieve and collect the expectation factor based on two data sources, air quality of 190 cities of past d days, and stock price of 62 selected A-share listed companies in China of past d days.

4.3.1 Air Quality. We analyze the correlation between air quality with iron ore price trend. As shown in Figure 7, the distribution of steel mills has an overlap with heavy air quality regions. In the map of China, we could observe that each province has been filled with different degree of different colors. The color reflects the degree of the mean air quality index. The size of light purple circles in the figure indicate the steel production volume statistics in 2016 calculated through top steels mills in China [47], shown in Figure 8. Note that, provinces without large-scale steel mills

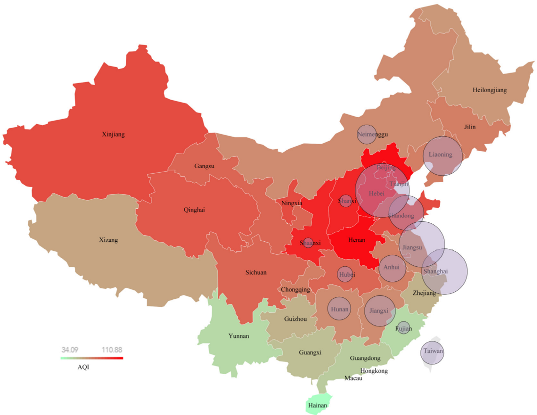


Fig. 7. Air quality heatmap and steels production distribution.[1]

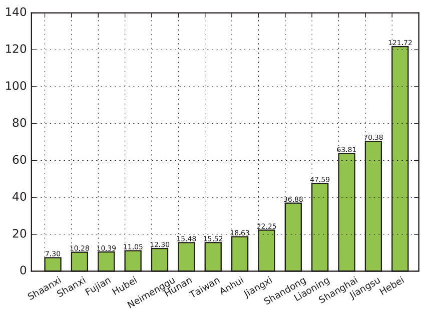
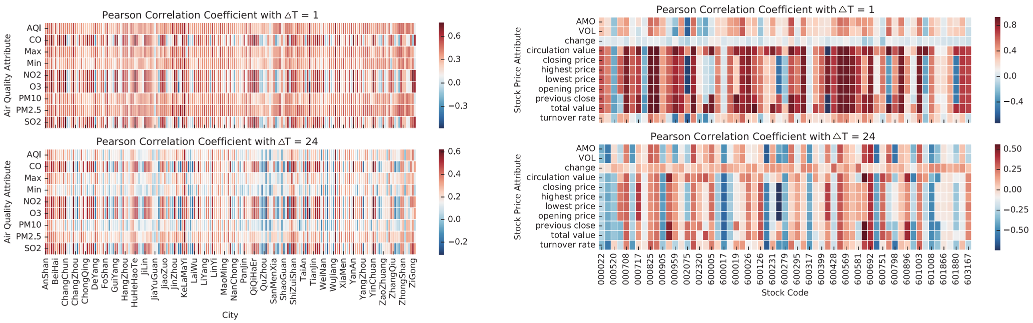


Fig. 8. Top steel production provinces in China.[1]



(a) Air Quality

(b) Stock Price

Fig. 9. Pearson correlation coefficient heatmap when $\Delta T = 1$ and 24, respectively.

also have dark red indicator in air quality index, such as Xinjiang, Qinghai. The reason behind may lie in the frequent pollutions of sand and dust, the common usage of coal for warming. We also should notice that there are some outliers with larger steel production and better air quality, e.g., Shanghai. We would discuss the possible reasons using an example of Shanghai. Shanghai has a big steel company *Baosteel*, one of the biggest iron and steel companies worldwide with a large amount of steel production [9]. Meanwhile, the geographic features (e.g., climatic condition) of Shanghai and the advanced air quality governance measures help the air circulation and further improve the air quality. Therefore, we should pay attention to extract useful features based on these factors for future price trend prediction. Thus, we confirm that air quality has a relationship with steel production, linking to iron ore usage and reflecting the expectation of short future. We further analyze the correlation of each air quality attribute of each city with IODEX trend, as presented in Figure 9(a). We observe that all attributes have correlations with iron ore price when $\Delta T = 1$. When $\Delta T = 24$, air pollutants (e.g., SO2, NO2, CO) keep strong correlation with the price, while the other air quality attributes lessen their impact on the future price. Even for one specific air quality attribute, the correlation with the price in different ΔT differs. Note that the

[1] Statistics from https://www.worldsteel.org.

corresponding spectrum of air pollutant CO in the bottom part is lighter than its counterpart in the upper part at most points.

4.3.2 Stock Price. We also analyze the correlation between stock price with iron ore price trend. Similar to air quality, stock price also has several components resulting in multiple attributes, e.g., daily transaction Volume (VOL) and daily transaction Amount (AMO). We conduct the analysis between each stock price attribute of each public A-listed company and iron ore price, and show the computation results in a heatmap in Figure 9(b). As shown in the upper part of the figure, we observe that stock price has strong correlation with iron ore price when $\Delta T = 1$. Circulation value, closing price, highest price, lowest price, opening price, previous close price, total value, all have distinct colors in their spectrums, indicating high positive impact up to 0.8 and negative impact to -0.4, respectively. Meanwhile, there are also some stock price features have weak correlation with the price, such as change. In the bottom part of the figure, we can see that there do not have any spectrum with strong correlation of future T business days. The correlation coefficient ranges of stock price features in all these companies is narrowed when $\Delta T = 24$, compared with the situation when $\Delta T = 1$. Most stock price features have impact on iron ore price with correlation coefficient locate in $[0.25, 0.5]$ and $[-0.5, -0.25]$. In addition, the stock price features in different ΔTs have different impacts, some increase (e.g., change) and some decrease (e.g., circulation value).

5 PRICE TREND PREDICTION

Due to the sequential and continuous characteristics of the price trend prediction in different future time intervals, we exploit classification-based approaches to solve this problem. In particular, we propose a hybrid model for **Pri**ce Tren**d** Pr**e**diction (Pride), which consists of multiple prediction models, i.e., Adaboost [13], SVM [10, 20], Naive Bayes [11, 27], and GBDT [14, 18], for different future time intervals. Specifically, we first select different relevant features from combining factors corresponding to different prediction time intervals, and then feed these selected features to form different training datasets. Based on these pairs of training datasets and objective datasets, we apply multiple independent predictors for each pair dataset, respectively. Note that all features need to be mapped and scaled to fit the models.

5.1 Feature Selection

We employ the randomized lasso method [25] to select important features for the three-class prediction problem. The reason why we select this method is because it can be beneficial for relevant features selection while avoiding overfitting. In our problem, given collected data matrix $X_{m \times n}$ and label matrix Y_m^T, with a prediction time interval ΔT, we first get a data matrix and define it as $X_{m' \times n}^{(\Delta T)}$ of the three non-overlapping categories, with each column X_k indicating a feature, $k = 1, 2, \ldots, n$, and a corresponding matrix $Y_{m'}^{(\Delta T)}$, where m' denotes the constrained number of data samples. We then adopt the randomized lasso approach to select relevant features.

When the lasso method chooses parameter λ as penalty for L1-norm of β_k, the randomized lasso considers changing the λ to a randomly selected value in the interval from λ to λ/α. In the parameter optimization procedure, the k-th column of matrix X would be rescale with a random parameter W_k in interval $[\aleph, 1]$. In this way, each type feature can be assigned a non-negative score to represent its relevancy. We select top s features to form a new feature matrix $X_{m' \times s}^{(\Delta T)}$, and then we form a data matrix $Z_{m' \times (s+1)}^{(\Delta T)}$ composed of $X_{m' \times n}$ and $Y_{m'}^T$ for models.

ALGORITHM 1: Pride Algorithm

```
Input: Feature Matrix $X_{m\times n}$, label vector $Y_m^T$, randomized parameter $\alpha$, selection parameter $s$, predictors
    set PRED = [Adaboost, SVM, GBDT, NB]
Output: predicted label vector $Y$
for $\Delta T = 1$ to $T$ do
    form $X_{m'\times n}^{(\Delta T)}$ and $Y_{m'}^{(\Delta T)}$
    set random variable $W_k \in [\alpha,1]$
    $\beta_{opt} := \arg\min_\beta \; \|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{n} \frac{|\beta_k|}{W_k}$
    list $\beta_{opt}$ in descend order, as $\beta_{opt}^{(1)}, \beta_{opt}^{(2)}, \ldots, \beta_{opt}^{(T)}$
    select top-s in $\beta_{opt}$ and reset the order number RS = $[rs_1, \ldots, rs_s]$
    $X_{m'\times s}^{(\Delta T)} \leftarrow$ features selected from $X_{m'\times n}^{(\Delta T)}$ with RS
    $Z_{m'\times(s+1)}^{(\Delta T)} \leftarrow (X_{m'\times n}^{(\Delta T)}, (Y_{m'}^{(\Delta T)})^T)$
    for i = 1 to 4 do
        while cross-validation do
            shuttle $Z^{tr}$ and $Z^{te}$, apply PRED(i) for $Z^{tr}$, apply best_paras(i) for $Z^{te}$
        end while
        obtain best_prediction(i)
    end for
    $i_{opt} :=$ arg max$_{ß}$ best_prediction(i)
    $Y_{\Delta T} \leftarrow$ best_prediction($i_{opt}$)
    best_pred($\Delta T$) $\leftarrow$ PRED($i_{opt}$) with best_paras($i_{opt}$)
end for
$Y \leftarrow$ concatenate $Y_{\Delta T}$, $\Delta T = 1$ to $T$
return predicted $Y^T$
```

5.2 Hybrid Prediction Classification Model

We propose a hybrid model consisting of multiple predictors, Adaboost, SVM, Naive Bayes, and GBDT [14, 18]. Adaboost is an algorithm using weak learners to create a strong predictor for classifications. It can avoid overfitting problem effectively while obtaining high accuracy. SVM is a strong discriminative predictor to generate an optimal separating hyperplane to distinguish data with different labels. It can cope with nonlinear and high-dimensionality problems effectively. Both Adaboost and SVM treat multi-class problems as a combination of several binary classification problems, and they decompose one-to-one strategy to address these issues. Naive Bayes (NB), a probabilistic predictor based on Bayes' theorem, is able to generate the probabilities of predicting instances, and then assign corresponding class labels for the instances. This predictor can perform well in small-scale datasets, and is robust to missing data. GBDT, an ensemble in decision tree style of weak predictors, produces a prediction result based on prediction results of all involved predictors. It is robust with outliers. The construction of the prediction model is in a stage-wise way, e.g., boosting approaches. Based on their advantages, we choose these four classification models.

We then apply this model into several data matrices for corresponding prediction time intervals separately, to achieve a sequential multi-class prediction of following T business days. In particular, given the input matrix $Z^{(\Delta T)} \in \mathbb{R}^{m'\times(s+1)}$ representing the price trend prediction of future ΔT, we attempt to apply each classification model, to discover the patterns from feature matrix X to the label matrix Y with three labels. Here, we should notice that the performance of a classification algorithm usually depends on which classification task it applied. For example, an algorithm good at the face recognition task may not achieve a good performance in the visual object categorization

Table 4. Description on Datasets

Datasets		Periods	Description
Price		12/01/2010–5/30/2017	1,632 records
Supply Factor	Inventory	5/8/2009–4/14/2017	11,970 records; 30 ports
	Import	1/31/2014–4/14/2017	978 records; 6 ports
Demand Factor	BDI	4/1/2016–5/5/2017	274 records
	Arrival Vessels	1/31/2014–4/14/2017	978 records; 6 ports
	Vessels Types	2016	11,000 records; 4 types
	GPS Trajectory	4/1/2016–4/23/2017	130,178,664 records
Expectation Factor	Air Quality	1/1/2014–4/20/2017	1,603,980 records; 190 cities, 7 types
	Stock Quotation	1/1/1993–5/5/2017	3,268,206 records; 62 companies, 9 types

task. The No Free Lunch theorem points out that NO algorithm will perform BETTER than all others when averaged over all possible problems [44–46]. That is, there is NO classification algorithm can be universally good [23]. In our study, it could be considered as a different classification task if the input is changed. In the price trend prediction of different ΔTs, after a few features are selected, we need to find which classification algorithms are performed well for these features. Specifically, some features play relatively short-term effects on the price trend prediction. For example, feature *inventory quantity in some ports* has a strong correlation with the price trend of the next day, as aforementioned. While, some features play relatively long-term effects on the price trend prediction. For example, feature *estimated arrival and departure rate* of vessels transporting bulk commodities, which have not yet arrived in China, can reflect the iron ore demand of purchasers in China, and then influence the price trend. Therefore, we propose a hybrid classification model composed of multiple classification algorithms, to solve different future time's prediction with appropriate classification algorithms to achieve best prediction performance.

6 EVALUATION

We evaluate the performance of our framework based on various real-world datasets from different domains. Iron ore is employed to conduct experiments, as an example of bulk commodities.

6.1 Experimental Settings

6.1.1 Datasets and Preprocessing. We collect iron ore price (IODEX) data and relevant influential factors data, and present them in Table 4. The description of these datasets in detail are as follows.

(1) Price data: We collect iron ore price (IODEX) data, from one of the leading steel-related information providers, qianzhan.com [35]. The dataset is provided every workday, and presented in U.S. dollar. After a data cleansing and preprocessing progress, we obtain 304 instances.
(2) Supply factor data: We collect data from two aspects, weekly iron ore inventory quantity and iron ore import quantity in 30 significant ports (e.g., port Qingdao and port Tianjin) and 6 significant ports (e.g., port Caofeidian and port Jingtang) of China, respectively, also from qianzhan.com [35].
(3) Demand factor data: We collect data from the perspective of vessels' arrival rate of 30 significant ports, GPS trajectory of seaborne vessels and BDI. The arrival rate data is also

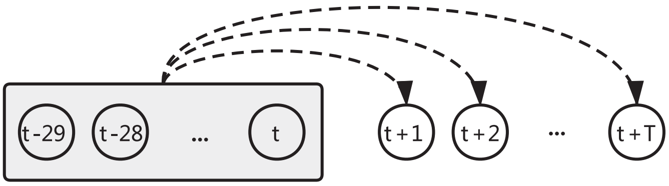


Fig. 10. Illustration of the price trend prediction.

retrieved from qianzhan.com [35]. For vessels GPS trajectory data, we first retrieve the four specific types vessels' data from Clarkson Research SIN [36], including name, DWT. Based on these data, we collect vessels' GPS trajectory data from hifleet.com[29], who provides free historical vessel trajectory retrieval with trial accounts. We also retrieve BDI value of each workday from Bloomberg Markets [24].

(4) Expectation factor data: We retrieve data from air quality in 190 cities in China and stock data of significant relevant enterprises. The daily air quality data is collected from aqistudy.cn [32], one of the predominant air quality information service providers. For stock data, we confirm that there are 62 listed companies in A-share market of China, including steel industry-related and overseas transportation-related companies. We then retrieve these historical data from NetEase Finance [12], who provides all historical data of all listed companies in China.

6.1.2 Evaluation Plan. We conduct price trend prediction tasks over upcoming T days simultaneously. Here, we set the threshold δ as 0.008 and T as 24. We use the data during the last 30 days for these T prediction tasks, as shown in Figure 10. For the price trend prediction of $(t + \Delta T)$'s day, the dataset is consisting of all involved data during last 30 days $[t - 29, t]$, and the price trend data of day $t + \Delta T$. $\Delta T \in [1,T]$. We then partition this dataset into non-overlapped training set and testing set by a ratio of 4:1. Specifically, we use the first 10 months data as the training data and the following 2.5 months data as the testing data.

6.1.3 Metrics and Ground Truth. We predict the changing trend of the price, and the ground truth can be obtained from its later readings. We define the following metrics to evaluate the prediction accuracy [57], to obtain the average precision, recall and F1-score of the three classes:

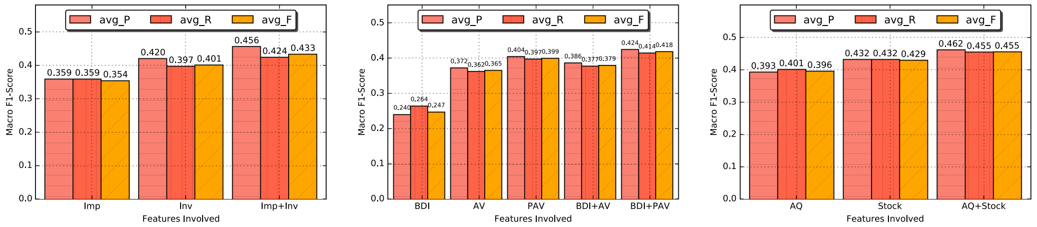
$$macro\_P = \frac{1}{n}\sum_{i=1}^{n} P_i, \tag{2}$$

$$macro\_R = \frac{1}{n}\sum_{i=1}^{n} R_i, \tag{3}$$

$$macro\_F1 = \frac{2 \times macro\_P \times macro\_R}{macro\_P + macro\_R}. \tag{4}$$

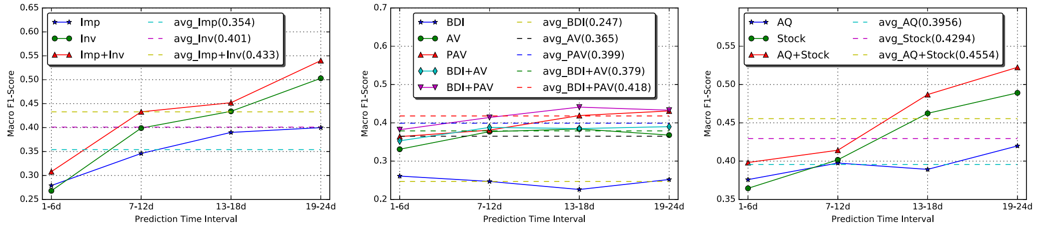
6.1.4 Baselines. We compare our method with following baselines. Note that each specific baseline method shares the same parameters in the price trend forecasting tasks on different ΔTs.

- Adaboost (Adaptive Boosting Algorithm): This baseline method inputs the distribution of training data into several individual predictors, and then strengthen their classification powers.
- GBDT (Gradient Boost Decision Tree): This baseline method is a powerful method for machine learning problems, with one tree constructed a time to fit the residual of preceded trees.



(a) supply (b) demand (c) expectation

Fig. 11. Average performance comparison of individual factors.



(a) supply (b) demand (c) expectation

Fig. 12. Duration-level performance comparison of individual factors.

- NB (Naive Bayes): This baseline method is a probabilistic predictor applying Bayes theorem for features with independence and objectives.
- SVM (Support Vector Machine): This baseline method is a typical predictor for multi-class problems. We utilize one-against-all strategy, and adopt LIBSVM here.
- Adaboost w/ fs: This baseline method first performs feature selection process and then applies the Adaboost algorithm for price trend prediction.
- GBDT w/ fs: Similarly, this baseline method selects relevant features and then uses GBDT to these features for the price trend prediction.
- NB w/ fs: Similarly, this baseline method uses feature selection first, and then applies NB to selected features to forecast the price trend of bulk commodities.
- SVM w/ fs: Similarly, this baseline method leverages SVM after feature selection process for forecasting the price trend of bulk commodities.

6.2 Results and Analysis

6.2.1 Study of Individual Factors. Experimental results of the study of individual factors and involved features, have been depicted in Figures 11 and 12. The continuous future 24 business days have been divided into four periods, 1–6 days (1–6d), 7–12 days (7–12d), 13–18 days (13–18d), and 19–24 days (19–24d). avg_P, avg_R, and avg_F denote average value of $macro\_P$, $macro\_R$, and $macro\_F1$ of these 24 days, respectively. For individual features, avg_Imp denotes average value of $macro\_F1$ of feature Imp. This definition is generalized, generating avg_Inv, $avg\_Imp + Inv$, and so on.

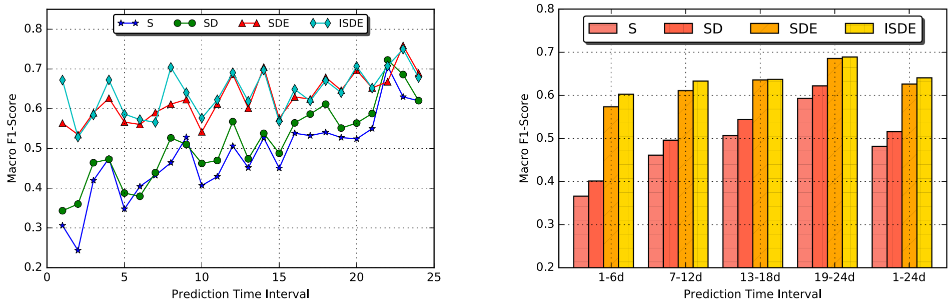
For the supply factor including inventory and import features, we should notice that the inventory quantity in each port can be influenced by the import quantity, since some trading companies warehouse some import bulk commodities in ports for future trades. These redundant features are correlated to some extent. To avoid a potential overfitting issue caused by the effect of multicollinearity, we apply Adaboost w/ fs process to evaluate the performance. From Figure 11(a), we observe that the combined factor obtains better performance than any individual feature (inventory *Inv* and import *Imp*) in the average value of *T* days' prediction, reflecting the effectiveness of these selected features. Besides, *Inv* plays more positive part than *Imp*. One reason may lie in that the larger number of participant ports in inventory, while only six ports can provide weekly import quantity due to limitation and hard-access of public open data. We also observe that all features and the factor can have better performance with increasing prediction time intervals, and achieve the best performance in the last prediction duration (i.e., 19–24d) in Figure 12(a). This observation indicates that participated features have delayed impact on the price, which is consistent with previous analyses.

For the demand factor, it includes seaborne cost, and estimated arrival and departure rate of vessels. The latter features reflect further latent demand of iron ore, and cannot be confirmed a linear correlation with the price trend directly. Therefore, we adopt SVM w/ fs for the performance evaluation, which solves nonlinear classification problems effectively. From Figure 11(b), we make the following observations. First, the surprising result is that *BDI* is not as effective as other features. One reason is that the impact from other aspects of iron ore price may play more important role than the seaborne fee change. Second, the combination of *BDI* and weekly arrival vessels *AV* can achieve better performance compared with *BDI* and *AV*. This observation illustrates the effectiveness and necessity of both individual features. Third, we introduce vessel GPS trajectory data to extract daily arrival and departure rate in relevant ports to expand and complement *AV* to formulate *PAV*. About these two features, we observe that *PAV* perform better than *AV* during all future intervals, and in terms of all metrics. This observation validates the advantages of vessel GPS trajectory data. From Figure 12(b), we observe that this whole factor improves the performance in all prediction durations. It indicates that the two features can complement each other.

For the expectation factor, as aforementioned, there are lots of features with some noisy data, e.g., air quality data of Shanghai. These noisy data would influence the forecast performance. Thus, we employ GBDT w/ fs for the performance evaluation, to cope with these outliers and be robust. From Figure 11(c), we observe that air quality *AQ* is not as effective as stock price *Stock*, with lower average value in three metrics. The reason may lie in that stock price of typical predominant iron ore and steel-related companies can reflect the expectation of markets' participants, who would play a more heuristic and direct impact on the iron ore price bid. Moreover, combination with *AQ* and *Stock* can have the best performance, indicating that the two types features can complement each other, with one from the subjective reality (human choices) and one from the objective reality (physical environments). From Figure 12(c), we observe *Stock* keeps a positive growing rate with the time interval, and expands its difference with *AQ*, which is consistent with our previous analysis.

6.2.2 Study of Factors Combination. We study the factors combination by applying our proposed Pride algorithm. Experimental results of this study on the effectiveness of factors are shown in Figure 13. We add different factors gradually, i.e., Supply factor (S), Demand factor (D), Expectation factor (E), and Iron ore historical price (I), and observe improvement on the average macro_F1 in the prediction of following *T* days.

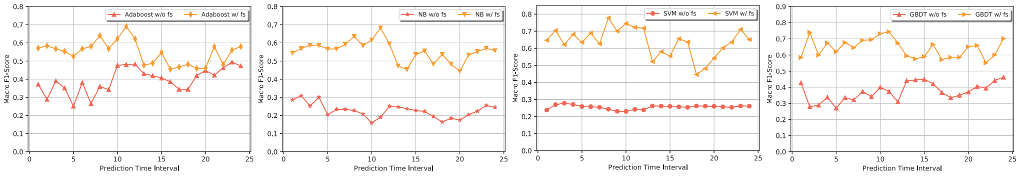
We first present the experimental results of day-level performance comparison in Figure 13(a). One observation is that SD can achieve better performance than S in almost every future day prediction, excepting predictions of two future days with slight differences. Meanwhile, the average macro_F1 value of SD is higher than S, showing almost 7% improvement. This verifies the effectiveness and justifies the necessity of the demand factor. Second, we observe that SDE and ISDE both have obvious improvements compared to S and SD in most time. Performance differences would decrease when the prediction time interval grows, and the biggest difference occurs during



(a) Day-level Performance

(b) Duration-level Performance

Fig. 13. Performance comparison of different factors.



(a) Adaboost

(b) NB

(c) SVM

(d) GBDT

Fig. 14. Performance comparison of different methods with and without feature selection process.

the future eighth days. One reason behind this may be that the stock price plays a more immediate role for the iron ore price change. The exception happens when the prediction time interval equals to 22, when the performance of S and SD both can surpass SDE and ISDE. Since we have selected features before feeding into prediction models, the possible reason may be that the historical inventory data and import data play significant roles for the future 22nd working day' prediction. We can see that ISDE obtains big improvements compared to SDE when the prediction time interval equals 1 and 8. One reason behind may be that the time-delay influence of the price is not constant, and has some changes. To examine the performance in a duration-level, we also present experimental results in Figure 13(b). One observation is that the performance of S is getting higher with time interval growing, and this observation also happens to SD, SDE, and ISDE. It indicates that the high influence of time-delay, and the macro_F1 score of S to ISDE achieve high value from 0.6 to 0.69 in 19–24d duration. Furthermore, we observe that the performance is growing in each time interval with the in turn factors adding, S, SD, SDE, ISDE, which validate the effectiveness of the factors added.

6.2.3 *Study of Feature Selection Process.*

The experimental results of baseline algorithms without feature selection are presented in Figure 14. Compared with the application of baseline algorithms with feature selection, the new experimental results of baseline algorithms without feature selection verify the effectiveness of the feature selection process. Each specific algorithm with feature selection process outperforms it without feature selection process in the respective figures. Moreover, we can observe that for some methods, e.g., NB and SVM, there has been notable performance difference between the method with and without feature selection process. While, for some ensemble methods, e.g., Adaboost and GBDT, there has been less performance difference between it with and without feature selection process. This may be because, without an effective feature selection process, there may have been redundant information and multi-collinearity between features. In this way, the ensemble approaches are able to cope with these situation and obtain good

Table 5. Comparison with Baselines w.r.t. Average macro_F1

ΔT clf	1	2	3	4	5	6	7	8	9	10	11	12
Adaboost w/ fs	0.477	0.488	0.547	0.455	0.467	0.482	0.459	0.461	0.579	0.481	0.56	0.58
GBDT w/ fs	0.473	0.453	0.536	0.556	0.483	0.537	0.483	0.444	0.533	0.552	0.57	0.557
NB w/ fs	0.522	0.580	0.555	0.656	0.637	0.446	0.482	0.543	0.602	0.636	0.71	0.651
SVM w/ fs	0.595	0.576	0.590	0.665	0.571	0.585	0.586	0.651	0.659	0.551	0.60	0.702
Pride (predictor)	**0.672** (SVM)	**0.596** (SVM)	**0.628** (SVM)	**0.678** (SVM)	**0.665** (NB)	**0.585** (SVM)	**0.656** (SVM)	**0.702** (SVM)	**0.659** (SVM)	**0.671** (NB)	**0.711** (NB)	**0.702** (SVM)
ΔT clf	13	14	15	16	17	18	19	20	21	22	23	24
Adaboost w/ fs	0.571	0.584	0.567	0.553	0.526	0.567	0.581	0.64	0.568	0.623	0.691	0.622
GBDT w/ fs	0.544	0.569	0.587	0.586	0.566	0.567	0.592	0.636	0.585	0.616	0.683	0.595
NB w/ fs	0.647	0.704	0.621	0.682	0.635	0.689	0.626	0.778	0.700	0.745	0.722	0.718
SVM w/ fs	0.584	0.740	0.599	0.675	0.620	0.677	0.645	0.691	0.695	0.732	0.744	0.674
Pride (predictor)	**0.674** (NB)	**0.753** (SVM)	**0.665** (SVM)	**0.720** (SVM)	**0.664** (SVM)	**0.762** (NB)	**0.698** (SVM)	**0.778** (NB)	**0.760** (NB)	**0.774** (SVM)	**0.826** (NB)	**0.735** (NB)

performance. However, NB assumes all features are independent and unrelated to each other. The possible multi-collinearity in features would impact NB's performance. For SVM, this method is hard to obtain good performance with the situation of many irrelevant features [43].

6.2.4 Comparison of Different Predictors. We compare the effectiveness of state-of-the-art predictors, i.e., Adaboost, GBDT, NB, SVM, and our Pride. The experimental results are presented in Table 5. Please note that the random choice method for this three-class prediction problem should be 33% for each day's prediction. We first observe that method NB and method SVM can obtain better performance than method Adaboost and method GBDT with higher average macro_F1, in most days' prediction. It helps present the effectiveness of the feature selection methods *randomized lasso* in this problem, which can select different useful features for the price trend prediction on different future days. It also indicates that within the selected features, there are lots of nonlinear features to improve the prediction performance. Moreover, we observe that the performance of price trend prediction on different ΔT is varying with one particular prediction method. More specifically, the price trend prediction on former several days often obtain worse performance with lower average macro_F1 values, compared with the prediction on latter several days, which usually get higher values. For example, note that the overall prediction performance of the last 5 days is much better than the performance of previous days. It validates our aforementioned analysis and effectiveness of introduced features, especially the estimated arrival and departure rate of seaborne vessels. Furthermore, one surprising observation is that the best performance of each method all occurs in the 23rd day, highlighting a best prediction time for iron ore-related entities. Note that the iron ore price is only published in working days. Here, the prediction of future 23rd day means predicting the price of the next 23rd working day, which may be the next 29th normal day (adding an extra 2 days each week). One possible reason of this surprising observation may be the estimated arrival rate, since seaborne vessels would reach coastal ports in China 15–30 days normally. In addition, we also present the specific algorithms selected for the price trend prediction on each ΔT, and demonstrate it as an additional row in Table 5. From the table, we can observe that the selected algorithm does not correspond to the baseline algorithm with best performance on the specific ΔT. In particular, the application of one specific baseline algorithm on each ΔT share the same parameters. Moreover, we can also observe that most of the selected algorithms

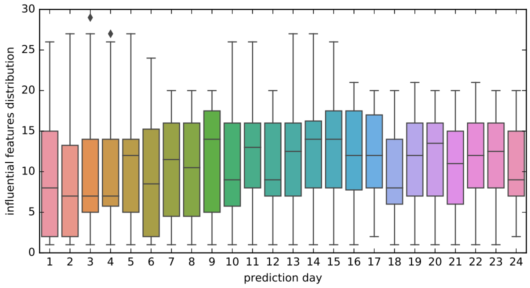


Fig. 15. Impact of time delay.

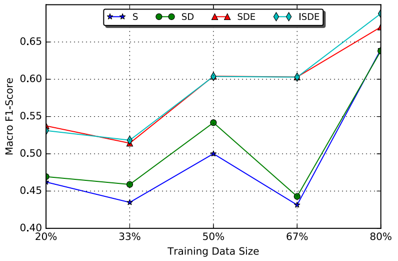


Fig. 16. Impact of training data size.

are SVM and NB. One reason behind may be the feature selection part. After the feature selection process, all selected features fed into the prediction models may eliminate many redundant information, and remain some nonlinear features.

6.2.5 Impact of Different Time Delay. The impact of different time-delay on each day's iron ore price trend prediction is presented in Figure 15, through identifying major influential features of each prediction. From the figure, we observe that most influential features are concentrated within a duration from past 17th to past 2nd day. It indicates that the data of all features aforementioned of past 2nd–17th days plays a positive role to predict the future ΔT days' iron ore price trend. Furthermore, we notice that the Inter-Quartile Range does not change much with different prediction days excepting a few bars, e.g., when ΔT equals to 1, 6, 7, 8, and 9. It refers to the concentration trend of most influential features. In addition, we also observe that the data from past 30th to past 20th day plays a certain role for the future iron ore price prediction, especially when ΔT is from 1 to 15. In this way, when we predict the iron ore price of future 15th day from Tth day (e.g., $T + 15$th), we need to feed data of past 30 days $[T - 30, T]$. All data is useful for the prediction. Meanwhile, when we predict the iron ore price of future 24th day (e.g., $T + 24$), feeding data from past 20 days $[T - 20, T]$ is enough and can obtain acceptable prediction results.

6.2.6 Impact of Training Data Size. Experimental results of impact of training data size are shown in Figure 16. We observe that the performance using different size of training data varies. When using factors *SDE* and *ISDE*, we find that the size of training data has a large impact on the macro F1-score, validating the significance of our feature selection process. Most relevant features are selected and form the training dataset, with less relevant features discarded. feature selection is critical in our problem, due to the imbalance between size of samples and size of features. Without the feature selection, size of features is much larger than the size of samples, making our model vulnerable and unstable. We further observe that *ISDE* has slight improvement compared with *SDE*. This improvement can be identified clearly when the size of training data grows to 80%. It indicates that the feature of the price plays a significant role on the future price trend prediction.

7 DISCUSSION

Having presented abundant research results, we will conduct a deep discussion about our work and several limitations from the following perspectives.

From the perspective of data, due to the data access difficulties, we are unable to obtain sufficient data. For instance, there exist daily statistic data of inventory and import in ports. However, in our work, we can only collect weekly inventory data and import data. If the data can be refined into daily data, then our proposed framework may be able to produce a more fine-grained prediction.

From the perspective of problems, we transform our price trend prediction problem into a T-length sequence of continuous sub-problems. It is a meaningful short-term price trend problem. Besides, a price trend prediction with longer term, such as several months, also plays an important role for markets participants and government administrators. In the future, we expect to extend our work to a farther future prediction, to achieve acceptable predictions with 3–6 months in advance.

From the perspective of implications, we provide a novel way for researchers to study the economic-related problem with cross-domain open data, to connect the physical world with the economic virtual world. With more open data, we may explore more cross-domain studies, which cannot be solved well with knowledge from one specific domain.

From the perspective of application to other bulk commodities, it is practical to apply our framework to other bulk commodities, e.g., soybean, copper, and so on. Here, we take soybean as an example. Similar with iron ore, the dynamically changing spot price of soybean is significantly affected by the balance between supply and demand, as well as some latent influential factors (named as expectation factor). For the supply factor, we adopt three sources providing soybean, i.e., soybean inventory in ports owned by trading companies, soybean import from overseas companies, and domestic soybean sold in markets. For the demand factor, it is difficult to openly obtain explicit demand from purchasers due to trade secrets. Similar to the iron ore example, we analyze the demand factor from the viewpoint of vessel transportation. The rationale behind is that the large volume of soybean exports transported through large-size vessels can reflect the high demand of soybean in the destination countries. For the expectation factor, we identify two important features according to our prior knowledge and previous studies [39], i.e., political risk and environmental impact. When the political risks increase, a series of price factors may change substantially. The environmental factors, e.g., weather conditions, also bring indirect effect on soybean price, by affecting the soybean production, transportation, and then finally the soybean supply-demand balance. In summary, through this application case, we demonstrate that it has potential to apply our framework to other bulk commodities for a short-term price trend prediction.

8 CONCLUSION AND FUTURE WORK

Prediction on price trend of bulk commodities has significant meanings for economic development worldwide, specifically for market participants and government administrators, to schedule plans, adjust policies, save costs, and increase profits. In this article, we employ various cross-domain open data to forecast the price trend of bulk commodities over multiple future days. The price trend prediction problem is transformed to a three-class prediction problem: rise, slight-change, and fall. We select the iron ore as an example of bulk commodities to present our analysis results and conduct experiments to demonstrate the effectiveness and efficiency of our proposed method. We identify three factors based on prior knowledge and previous studies, i.e., supply factor, demand factor, and expectation factor. We then categorize all collected open data into the three factors and validate their correlations with the price. Relevant features are selected and fed into a proposed hybrid classification model, to achieve a sequentially three-class prediction of continuous T business days. Finally, we conduct extensive experiments to evaluate the performance of our framework using nine real-world cross-domain open datasets. Experimental results show that our method achieves surpassing performance compared with state-of-the-art baselines.

In the future, we plan to broaden and deepen our work from two directions. First, we intend to study a longer term price trend prediction for bulk commodities, with more cross-domain open data. Second, we plan to introduce deep neural network-based models for time-series data, to study the application of these models in coarse-grained and insufficient economic-related data.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers and the editors for their valuable comments and helpful suggestions.

REFERENCES

[1] BDI. 2017. Baltic Dry Index. Retrieved from http://www.balticexchange.com.

[2] Chao Chen, Shuhai Jiao, Shu Zhang, Weichen Liu, Liang Feng, and Yasha Wang. 2018. TripImputor: Real-time imputing taxi trip purpose leveraging multi-sourced urban data. *IEEE Trans. Intell. Transport. Syst.* 99 (2018), 1–13.

[3] Chao Chen, Daqing Zhang, Xiaojuan Ma, Bin Guo, Leye Wang, Yasha Wang, and Edwin Sha. 2016. Crowddeliver: Planning city-wide package delivery paths leveraging the crowd of taxis. *IEEE Trans. Intell. Transport. Syst.* 18, 6 (2016), 1478–1496.

[4] Longbiao Chen, Daqing Zhang, Xiaojuan Ma, Leye Wang, Shijian Li, Zhaohui Wu, and Gang Pan. 2016. Container port performance measurement and comparison leveraging ship GPS traces and maritime open data. *IEEE Trans. Intell. Transport. Syst.* 17, 5 (2016), 1227–1242.

[5] Longbiao Chen, Daqing Zhang, Gang Pan, Xiaojuan Ma, Dingqi Yang, Kostadin Kushlev, Wangsheng Zhang, and Shijian Li. 2015. Bike sharing station placement leveraging heterogeneous urban open data. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 571–575.

[6] Longbiao Chen, Daqing Zhang, Gang Pan, Leye Wang, Xiaojuan Ma, Chao Chen, and Shijian Li. 2014. Container throughput estimation leveraging ship GPS traces and open data. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 847–851.

[7] Longbiao Chen, Daqing Zhang, Leye Wang, Dingqi Yang, Xiaojuan Ma, Shijian Li, Zhaohui Wu, Gang Pan, Thi-Mai-Trang Nguyen, and Jeremie Jakubowicz. 2016. Dynamic cluster-based over-demand prediction in bike sharing systems. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 841–852.

[8] Javier Contreras, Rosario Espinola, Francisco J. Nogales, and Antonio J. Conejo. 2003. ARIMA models to predict next-day electricity prices. *IEEE Trans. Power Syst.* 18, 3 (2003), 1014–1020.

[9] Shanghai Baosteel Group Corporation. 2019. Retrieved from http://www.baosteel.com.

[10] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.* 20, 3 (1995), 273–297.

[11] Pedro Domingos and Michael Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.* 29 (1997), 103–130.

[12] NetEase Finance. 2017. Quotations. Retrieved from http://quotes.money.163.com.

[13] Yoav Freund and Robert E. Schapire. 1997. A desicion-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55, 1 (1997), 119–139.

[14] Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29, 5 (2001), 1189–1232.

[15] Bin Fu, Jialiu Lin, Lei Li, Christos Faloutsos, Jason Hong, and Norman Sadeh. 2013. Why people hate your app: Making sense of user feedback in a mobile app store. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1276–1284.

[16] Yanjie Fu, Yong Ge, Yu Zheng, Zijun Yao, Yanchi Liu, Hui Xiong, and Jing Yuan. 2014. Sparse real estate ranking with online user reviews and offline moving behaviors. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE, 120–129.

[17] B. Guo, Y. Liu, W. Wu, Z. Yu, and Q. Han. 2017. ActiveCrowd: A framework for optimized multitask allocation in mobile crowdsensing systems. *IEEE Trans. Hum.-Mach. Syst.* 47, 3 (June 2017), 392–403.

[18] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2008. *The Elements of Statistical Learning*. Springer.

[19] Stefan Hellmer and Joakim Ekstrand. 2013. The iron ore world market in the early twenty-first century-the impact of the increasing Chinese dominance. *Mineral Econ.* 25, 2–3 (2013), 89–95.

[20] Chih-Wei Hsu and Chih-Jen Lin. 2002. A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Netw.* 13 (2002), 415–425.

[21] Weimin Ma, Xiaoxi Zhu, and Miaomiao Wang. 2013. Forecasting iron ore import and consumption of China using grey model optimized by particle swarm optimization algorithm. *Resources Policy* 38, 4 (2013), 613–620.

[22] Yiqun Ma. 2013. Iron ore spot price volatility and change in forward pricing mechanism. *Resources Policy* 38, 4 (2013), 621–627.

[23] Malik Magdon-Ismail. 2000. No free lunch for noise prediction. *Neural Comput.* 12, 3 (2000), 547–564.

[24] Bloomberg Markets. 2017. BDIY Quote—Baltic Dry Index. Retrieved from https://www.bloomberg.com/quote/BDIY:IND.

[25] Nicolai Meinshausen and Peter Bühlmann. 2010. Stability selection. *J. Roy. Stat. Soc.: Ser. B (Stat. Methodol.)* 72, 4 (2010), 417–473.

[26] Maria Isabel Wolf Motta Morandi, Luis Henrique Rodrigues, Daniel Pacheco Lacerda, and Isaac Pergher. 2014. Foreseeing iron ore prices using system thinking and scenario planning. *Syst. Pract. Action Res.* 27, 3 (2014), 287–306.

[27] Andrew Y. Ng and Michael I. Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems* (2002), 841–848.

[28] National Bureau of Statistics. 2017. Retrieved from http://www.stats.gov.cn/.

[29] Fleet Online. 2017. Ship dataset. Retrieved from http://www.hifleet.com.

[30] OpenSea. 2018. Retrieved from https://opensea.pro/.

[31] Y. Ouyang, B. Guo, X. Lu, Q. Han, T. Guo, and Z. Yu. 2018. CompetitiveBike: Competitive analysis and popularity prediction of bike-sharing apps using multi-source data. *IEEE Trans. Mobile Comput.* (2018), 1–1. DOI: https://doi.org/10.1109/TMC.2018.2868933

[32] Online Air Quality Monitoring Platform. 2017. Air Quality. Retrieved from https://www.aqistudy.cn.

[33] Platts. 2017. Platts Iron Iron Index. Retrieved from http://platts.com.

[34] Alexander Pustov, Alexander Malanichev, and Ilya Khobotilov. 2013. Long-term iron ore price modeling: Marginal costs vs. incentive price. *Resources Policy* 38, 4 (2013), 558–567.

[35] Qianzhan. 2017. IRON ORE. Retrieved from http://d.qianzhan.com/xdata.

[36] Clarkson Research. 2017. Shipping Intelligence Network. Retrieved from https://sin.clarksons.net.

[37] Jingbo Shang, Yu Zheng, Wenzhu Tong, Eric Chang, and Yong Yu. 2014. Inferring gas consumption and pollution emission of vehicles throughout a city. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 1027–1036.

[38] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, and Ryosuke Shibasaki. 2016. Prediction and simulation of human mobility following natural disasters. *ACM Trans. Intell. Syst. Technol.* 8, 2 (2016), 29:1–29:23.

[39] SWS. 2018. SWS Research. Retrieved from http://www.swsresearch.com.

[40] Linda Wårell. 2014. The effect of a change in pricing regime on iron ore prices. *Resources Policy* 41 (2014), 16–22.

[41] Kun Wang, Hezhong Tian, Shenbing Hua, Chuanyong Zhu, Jiajia Gao, Yifeng Xue, Jiming Hao, Yong Wang, and Junrui Zhou. 2016. A comprehensive emission inventory of multiple air pollutants from iron and steel industry in China: Temporal trends and spatial variation characteristics. *Sci. Total Environ.* 559 (2016), 7–14.

[42] Yi-Hsien Wang. 2009. Nonlinear neural network forecasting model for stock index option price: Hybrid GJR–GARCH approach. *Expert Syst. Appl.* 36, 1 (2009), 564–570.

[43] Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. 2001. Feature selection for SVMs. In *Advances in Neural Information Processing Systems.* MIT Press, 668–674.

[44] David H. Wolpert. 1996. The lack of a priori distinctions between learning algorithms. *Neural Comput.* 8, 7 (1996), 1341–1390.

[45] David H. Wolpert, William G. Macready et al. 1995. *No Free Lunch Theorems for Search.* Technical Report. Technical Report SFI-TR-95-02-010, Santa Fe Institute.

[46] David H. Wolpert, William G. Macready et al. 1997. No free lunch theorems for optimization. *IEEE Trans. Evolution. Comput.* 1, 1 (1997), 67–82.

[47] WorldSteel. 2017. Steel Statistical Yearbook 2016. Retrieved from https://www.worldsteel.org.

[48] Kuang Xiao, Yuku Wang, Guang Wu, Bin Fu, and Yuanyuan Zhu. 2018. Spatiotemporal characteristics of air pollutants (PM10, PM2.5, SO2, NO2, O3, and CO) in the inland basin city of Chengdu, southwest China. *Atmosphere* 9, 2 (2018), 74.

[49] Anamika Yadav, Rajagopal Peesapati, and Niranjan Kumar. 2017. Electricity price forecasting and classification through wavelet-dynamic weighted PSO-FFNN approach. *IEEE Syst. J.* 12, 4 (2018), 3075–3084.

[50] Dongqing Zhang, Guangming Zang, Jing Li, Kaiping Ma, and Huan Liu. 2018. Prediction of soybean price in China using QR-RBF neural network model. *Comput. Electron. Agric.* 154 (2018), 10–17.

[51] Sha Zhao, Shijian Li, Julian Ramos, Zhiling Luo, Ziwen Jiang, Anind K. Dey, and Gang Pan. 2019. User profiling from their use of smartphone applications: A survey. *Pervas. Mobile Comput.* 59 (2019), 101052.

[52] Sha Zhao, Julian Ramos, Jianrong Tao, Ziwen Jiang, Shijian Li, Zhaohui Wu, Gang Pan, and Anind K. Dey. 2016. Discovering different kinds of smartphone users through their application usage behaviors. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM, 498–509.

[53] Sha Zhao, Yizhi Xu, Xiaojuan Ma, Ziwen Jiang, Zhiling Luo, Shijian Li, Laurence T. Yang, Anind K. Dey, and Gang Pan. 2019. Gender profiling from a single snapshot of apps installed on a smartphone: An empirical study. *IEEE Trans. Industr. Info.* (Early Access). (2019).

[54] Yu Zheng and Xing Xie. 2011. Learning travel recommendations from user-generated GPS traces. *ACM Trans. Intell. Syst. Technol.* 2, 1 (Jan. 2011).

[55] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. 2015. Forecasting fine-grained air quality based on big data. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Sydney, Australia, 2267–2276.

[56] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, 791–800.

[57] Zhihua Zhou. 2016. *Machine Learning*. TsingHua Press, Beijing.

Received May 2019; revised July 2019; accepted August 2019