Expert Systems With Applications 183 (2021) 115304

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A multi-stage data mining approach for liquid bulk cargo volume analysis based on bill of lading data

Suhyeon Kim [a,1], Wonho Sohn [a,1], Dongcheol Lim [a], Junghye Lee [b,*]

[a] Department of Industrial Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea
[b] Department of Industrial Engineering & Graduate School of Artificial Intelligence, UNIST, Ulsan 44919, Republic of Korea

ARTICLE INFO

Keywords:
Maritime logistics
Liquid bulk cargo volume
Data mining
Item segmentation
Exploratory volume analysis
Volume prediction

ABSTRACT

Liquid bulk cargo (LBC) volume analysis has received considerably great attention recently since LBC is a valuable and high-demand cargo. Thus, it is important to establish an analysis system for LBC volume, as it can help inform strategies for port planning and management. Nevertheless, LBC volume analysis is a challenging task for researchers because trends in LBC volume are highly volatile and non-stationary. In this paper, a new framework for enabling informative LBC volume analysis based on bill of lading (BL) data is proposed, which consists of three parts: item segmentation, exploratory volume analysis, and volume prediction. Firstly, an innovative item segmentation system using item texts of BL data was developed, which can generate subcategory as well as category information of LBC items that existing system cannot provide. Next, exploratory volume analysis was performed to understand the volume characteristics of each categorized and subcategorized item in terms of geography and timeline. Lastly, manifold learning- and deep learning-based time series techniques were proposed to increase LBC volume prediction accuracy compared with existing statistical models. The experimental results for volume prediction show the accuracy increased by 34% and 18% in average at category and subcategory levels over baseline models. It is believed that our proposed method will be helpful for stakeholders in maritime logistics, giving them the insights that they need to make better decisions.

1. Introduction

Maritime logistics is one of the most important sectors in global trade and supply chain networks (Fagerholt et al., 2017; Zhou et al., 2019). In maritime logistics, it is essential to establish initial plans relevant to port operation and management since such plans determine the mid- and long-term direction of ports. Port cargo volume analysis has become the basis for designing plans for port development as it provides various stakeholders, including shippers, consignees, port authorities, and government agents, with important information (Lee & Lee, 2016). To be specific, as cargo types are diverse and the demand for large-scale cargoes (i.e. bulk cargoes) is increasing worldwide, the size of cargo vessels is rapidly increasing in order to improve the efficiency of transportation in terms of time and cost (Merk et al., 2015). Hence, it is necessary to construct the port infrastructure that can handle these various bulk cargoes and related vessels. Port cargo volume analysis plays an important role in making optimal decisions about port size and utility, while also helping to manage vessel scheduling and providing a competitiveness index of ports (Lee, Song, Park, & Sohn, 2014).

In port cargo volume analysis, data for customs and data for a port community system (PCS) (here onwards, customs data and PCS data respectively) are used worldwide in the maritime industry (Adland et al., 2017; Guszczak & Mencarelli, 2020). Both data generally provide port cargo volume statistics aggregated based on bill of lading (BL) data, a detailed receipt of a shipment of goods and the standard for cargo import–export declaration in ports. That is, BL data is a micro-level source that can provide significant and precise information about cargoes, which is not revealed in the customs data and PCS data, such as item texts and ship names. For aggregating port cargoes volume, especially, 6-digit harmonized system (HS) codes are an important criterion of item segmentation, and the HS code-based item segmentation system (HSCS), which adopts the HS code as a criterion of categorization, can transform BL data into the categorized cargo volume data (Adland et al., 2017; Lee, 2020). Fig. 1 shows the process of the HSCS.

Most studies about port cargo volume analysis have primarily focused on containers, not liquids or gases (Kim, Oh, & Woo, 2018).

* Corresponding author.
E-mail addresses: suhyeonkim@unist.ac.kr (S. Kim), wonhosohn@unist.ac.kr (W. Sohn), dongclim@unist.ac.kr (D. Lim), junghyelee@unist.ac.kr (J. Lee).
[1] The authors equally contributed.

https://doi.org/10.1016/j.eswa.2021.115304
Received 15 August 2020; Received in revised form 27 November 2020; Accepted 27 May 2021
Available online 7 June 2021

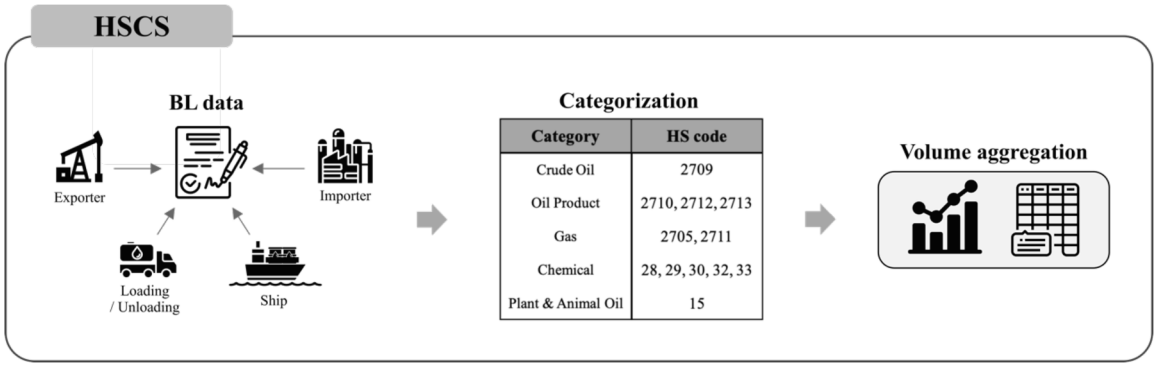


Fig. 1. Process of HSCS.

However, analyzing liquid bulk cargo (LBC) volume is a matter of urgency since interest in LBCs is expected to provide greater direct and indirect economic ripple effects in oil-related industries (e.g. automotive, shipbuilding, and petrochemical sectors) (Fjærtoft, 2015). In fact, LBCs, transferred to a large industrial complex, can create high added value; for example, 0.18 tons of naphtha at approximately $ 125 can be refined into petrochemical products, creating $ 9,000 added value (Kim & Ko, 2007). As LBCs have emerged as a competitive cargo, systematic volume analysis specific to LBCs is of great significance.

As previously mentioned, it is important to analyze LBC volume for LBC-related port management, and the use of BL data involving more specific information than the customs data and PCS data about LBC volume will be helpful for the analysis. In this study, a multi-stage framework that can provide comprehensive information about the volume of LBCs is developed by using BL data which enables the analysis to be performed at various and in-depth levels. Our framework uses a three-stage data mining approach for LBC volume analysis: item segmentation, exploratory volume analysis, and volume prediction. In the first stage, a new item segmentation system based on item texts in BL data is introduced. In the second stage, an exploratory volume analysis for LBCs is conducted from various points of view such as timeline and geography. In the last stage, manifold learning- and deep learning-based multivariate techniques are proposed for accurate LBC volume prediction to reflect the LBC properties by considering the external factors.

The paper is organized as follows. Section 2 presents background and previous studies relevant to this study. A description of the new framework follows in Section 3. Section 4 provides a case study containing data description, experimental design, and the comparison results with the existing methods. Section 5 and 6 describe the discussion and conclusion of this study, respectively.

2. Background

2.1. Related work for LBC volume analysis

Over the past few decades, great effort has been made in LBC volume analysis, ranging from exploratory analysis based on cargo volume statistics to cargo volume prediction. For LBC exploratory volume analysis, most of the studies are mainly focused on identifying the potential sources (cargoes or regions) for LBC trade or building the plans and strategies related to the operation of ports for LBCs by simply analyzing the statistics of LBC volume. Zhang and Xing (2018) analyzed the status of crude oil volume in global trading and investigated history of oil consumption in China and India by using statistical and geographical exploratory analysis methods. From this, they represented the change in the concentration degrees for Crude Oil importing sources of China and India from 2006 to 2015 and discovered the overlapping regions which can be important sources for Crude Oil trade between two countries. In addition, Wang et al. (2019) identified the intuitive information of LBCs in the 25 ports such as St. Petersburg and Amderma lying along the Northern Sea Route (NSR) based on several exploratory data analyses in terms of timeline and geographical views; the purpose of their study is to build port planning by considering situation and future demand for the ports related to the NSR.

Furthermore, for LBC volume prediction, current studies have conducted forecasting the LBC volume by using traditional time series models. Jai Sankar et al. (2016) used the export data of LBC from 1987 to 2015 collected from Chennai port in India and analyzed using an autoregressive integrated moving average (ARIMA) model to forecast LBC volume in 2020–21. The volume prediction of cargoes whose types were divided into three categories: container, liquid, and general cargoes, for the North Port of Busan in Korea in 2001–2007 was studied by Kim (2008), and LBC volume was predicted by using a seasonal ARIMA (SARIMA) model. Besides, Kim and Woo (2017) predicted the monthly LBC volume of Ulsan port in Korea by using moving average and regression models with seasonality. Further, Kim et al. (2018) shared the purpose of the study from Kim and Woo (2017), but they developed the two-way seasonality multiplied regressive model and compared it with the existing statistical models. However, the prediction models of afore-mentioned studies are applicable only to univariate time series data; it means they can just reflect the property of target variable without other related factors.

2.2. Motivations and objectives

LBC volume analysis based on BL data is more beneficial than the customs data and PCS data, but it is not an easy task due to the following problems related to two aspects: (1) the HSCS-based BL data and (2) properties of LBC volume itself.

(1) The HSCS also has two issues. Firstly, since the descriptions of the HS code are somewhat vague, the codes are notoriously difficult to use as informative item categories (or subcategories). Moreover, in BL data, many items are allocated the wrong HS code. When a carrier (or its agent) enters the HS codes into the BL data based on item texts, he/she may mistype the codes because of the ambiguity in code description. This results in serious errors in aggregating port cargo volume statistics, which means incorrect information is provided to stakeholders in maritime logistics.

(2) In general, since liquid cargoes are much bulkier than container cargoes, LBC volume has very irregular and highly volatile properties resulting from the industrial structure of bulk cargoes. LBC volume is also sensitive to external factors such as exchange rate, the economic situation, and the balance of payment. In addition, patterns and trends in the volume flow of not just each category but subcategory of LBCs are different. For example, Light Oil, Gasoline and Naphtha which all belong to the same category, Oil Products, may have different volumetric properties. Thus, in LBC volume analysis, it is necessary to consider the detailed properties of (sub)categorized LBC volume.

However, the existing LBC-related studies are not sufficient to tackle these problems. First, they focused on volume prediction at the category

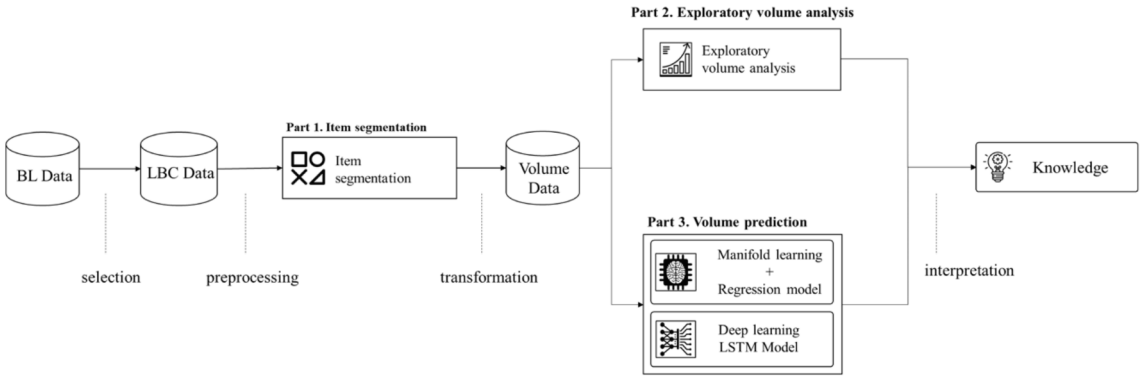


Fig. 2. Framework of LBC volume analysis.

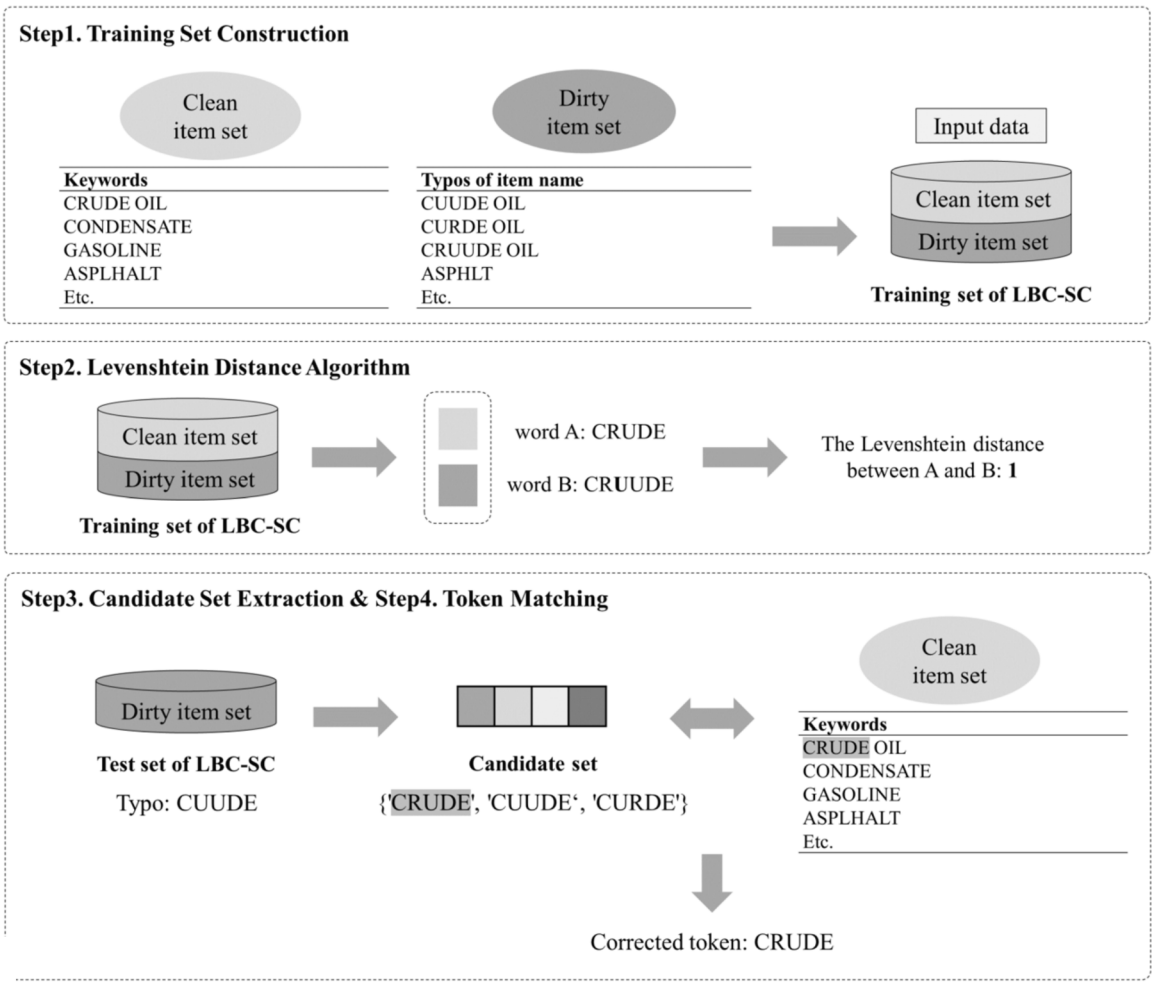


Fig. 3. Example of LBC-SC process.

level only. In addition, there is no study that takes into account data errors caused by HSCS. Furthermore, they mostly used traditional time series models, especially univariate time series models that cannot capture the relationships between multiple items. Lastly, there have been only a few studies that have considered external data, an important factor in LBC volume prediction.

A novel multi-stage data mining framework proposed in this study to resolve the abovementioned problems and improve existing studies consists of three parts and each one has its own contributions. First, item segmentation, which serves as the cornerstone of the next stages, is a new system that uses item texts in BL data instead of the HS code in categorizing cargo items, which provides accurate information by resolving the existing problems of HSCS and enables subcategory-level analysis. Second, exploratory analysis for LBC volume can uncover information which hidden in the subcategorized data. Lastly, the proposed multivariate machine learning and deep learning models for volume prediction are possible to reflect the LBC properties effectively by considering temporal information and multiple factors; it can lead to an accurate prediction for LBC volume. To the best of our knowledge, this is the first work to present an integrated decision support system for LBCs based on the BL data using multiple data mining techniques.

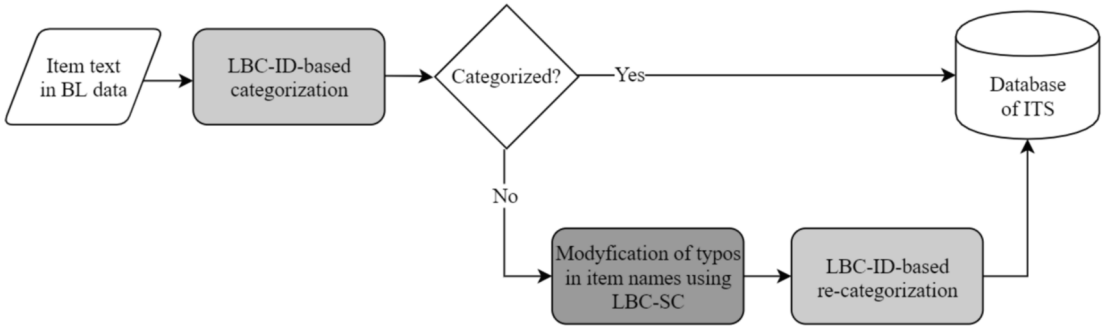


Fig. 4. Overall process of ITS.

3. Proposed framework

In this section, a new framework for LBC volume analysis is described, consisting of three parts: item segmentation, exploratory volume analysis, and volume prediction (Fig. 2). Details of each part are given in the subsections below.

3.1. Item segmentation

To resolve the aforementioned problems, an item segmentation system based on item text, called ITS, is proposed. The reason for using item text data is that it implies the actual meaning of cargo items and it can also help reduce the errors of the HS code caused by its ambiguous descriptions. The ITS contains two components: LBC-specific item dictionary (LBC-ID), and LBC-specific spell checker (LBC-SC). A detailed explanation of each is followed.

The LBC-ID, created for the ITS, is a domain-specific item dictionary which plays the role in criteria for item categorization and is used for the LBC-SC. The LBC-ID includes categories, subcategories, and keywords, and its structure is that the keywords belong to subcategory, and each subcategory belongs to category. The subcategories within each category of LBC items are defined by several criteria: chemical characteristics, materials or components, and industrial utilization. Then, the keywords are extracted from item texts, which are the criteria for item categorization; the keywords that belong to other subcategories are excluded from the LBC-ID. Based on the LBC-ID, each item is classified, whose text contains at least one of the main keywords, into the corresponding subcategory and category (i.e., item categorization).

In addition, since item texts in BL data are written in various ways by different consignors, they have many typos. This means a spell checker is essential for modifying typos correctly. However, general spell checkers using plain English dictionaries (Peterson, 1980) are not suitable for LBC-specific case. Therefore, the LBC-SC (**Algorithm 1**) is developed, consisting of four steps based on the LBC-ID and the Levenshtein distance algorithm, one of the most popular spell checkers (See Levenshtein (1966) for details).

Algorithm 1. LBC-SC

Inputs: LBC-ID, item texts of BL data, and q
Output: Corrected item texts
Step 1. Training set construction:
$C \leftarrow$ a clean item set containing q-times duplicated keywords of the LBC-ID
$D \leftarrow$ a dirty item set containing the typos of item texts
Generate a training set, $S = \{C, D\}$
Step 2. Levenshtein distance algorithm:
for all tokens $v \in S$ **do**
  for all tokens $j \in S$ $(j \neq v)$ **do**
    Calculate the minimum edit distance between v and j
    $edit(v,j)$ * Levenshtein minimum edit distance (v, j))
  end for
  Calculate the token probabilityp_v
  n_v: the frequency of token$v$
  n_S: the size of$S$

(*continued on next column*)

(*continued*)

Algorithm 1. LBC-SC

  $p_v * \frac{n_v}{n_S}$
  return $edit(v,j)$ andp_v
end for
for all typos $w \in D$ **do**
  Step 3. Candidate set extraction:
  A_w^1* a candidate set of w, which is$\{u \in S | edit(u,w) = 1\}$
  A_w^2* a candidate set of w, which is$\{u \in S | edit(u,w) = 2\}$
  Sort A_w^1, A_w^2 in the decreasing order of the token probability respectively
A_w* Union(A_w^1, A_w^2)
  Step 4. Token matching:
  for all $g \in A_w$ **do**
    if $g \in C$ **then**
      c_w*g
      break
    else
      c_w *w
    end if
  end for
  return c_w
end for

Specifically, Fig. 3 shows an example of the LBC-SC process. As shown in Algorithm 1 and Fig. 3, input data (indicated as a training set) contains tokens of the keywords in the LBC-ID (i.e. a clean set) and non-classified item texts in item categorization (i.e. a dirty set). Each token in the clean set is replicated at least q times to complement the low number of unique clean tokens since the number of tokens in the clean and dirty set in the training set should be balanced for adequate model learning. Next, the probability of each token is estimated, which is its frequency divided by the total number of tokens in the training set. The Levenshtein edit distance algorithm is trained based on all pairs of the training set to calculate the pairwise minimum edit distances between tokens. To achieve the goal of the LBC-SC, the tokens in the dirty set are used as a test set. When the test set is applied to the trained model, a candidate set is extracted based on the minimum edit distances and the probabilities of the tokens. For example, the candidate set for 'Cuude' as a typo of 'Crude' might contain 'Crude', 'Curde', and 'Cruda' according to the minimum edit distance and the probability of each token. For token matching step, incorrect tokens in the test set are changed into correct tokens if each token in the candidate set matches any token in the clean set. If several tokens in the candidate set belong to the clean set, the token with the highest probability and minimum edit distance is returned as the correct token.

Finally, item texts corrected by the LBC-SC are re-classified into subcategories and categories based on the LBC-ID. The typos in item texts are gradually reduced by updating new keywords to the LBC-ID and repeating the LBC-SC. In summary, through the ITS, it is possible to generate correctly categorized and subcategorized LBC volume data which can be used in exploratory volume analysis and volume prediction. The illustration for overall process of the ITS is shown in Fig. 4.

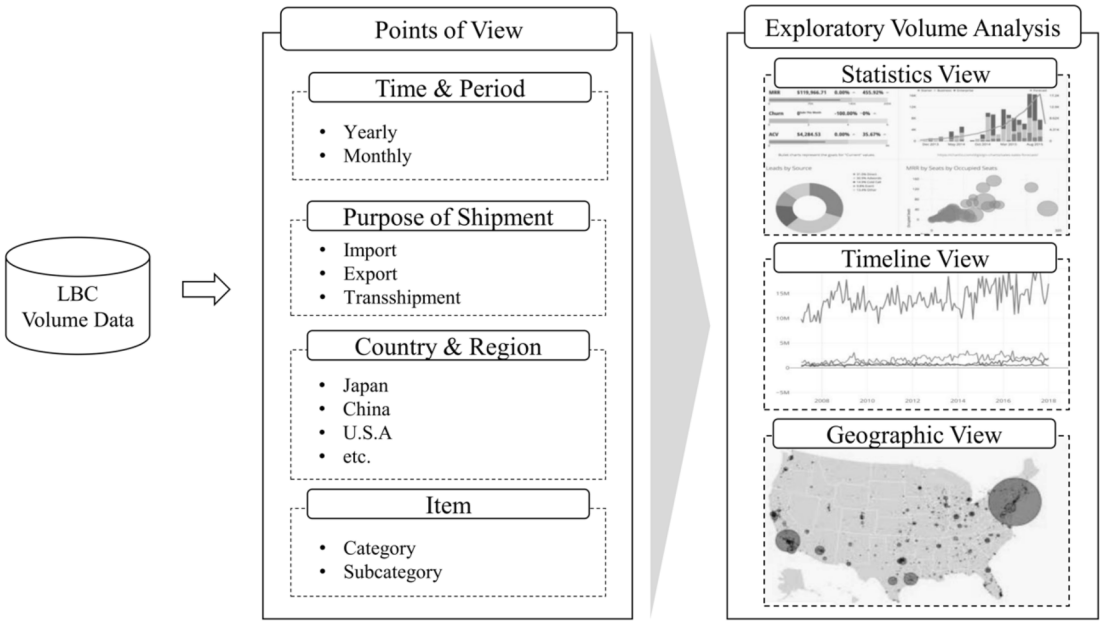


Fig. 5. Exploratory volume analysis for LBCs.

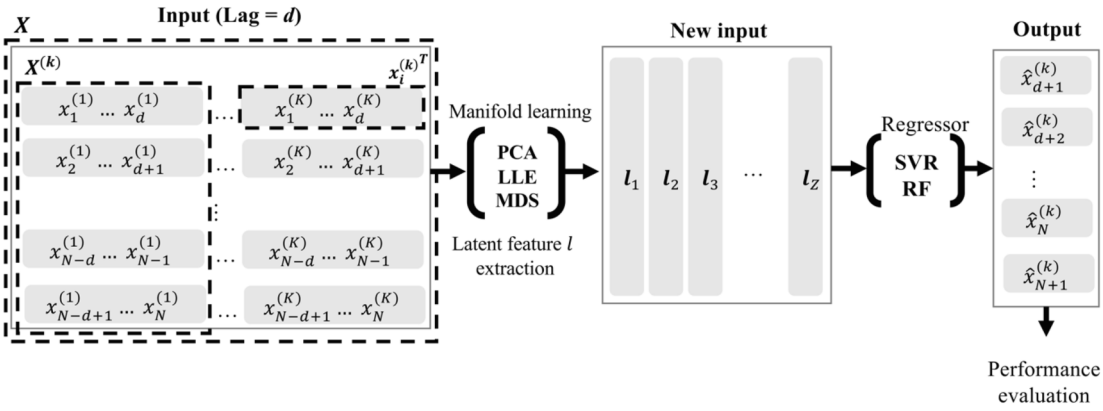


Fig. 6. Manifold learning-based model for LBC volume prediction.

3.2. Exploratory volume analysis

Exploratory analysis for data is a data analysis approach that summarizes the characteristics of the data, often along with visual methods (Yu, 1977). In this section, to carry out exploratory volume analysis for LBCs is proposed by aggregating descriptive statistics of LBC volume by considering combinations of various points of view such as items in both category and subcategory levels, countries, period, and the purpose of shipment categorized by import, export, and transshipment (Spyridoula, 2019). Exploratory volume analysis can provide intuitive information about the flow of port cargo volume and new insights which cannot be seen in raw data. Fig. 5 shows an exploratory volume analysis process for LBCs.

3.3. Volume prediction

Typical univariate time series models widely used in port cargo volume prediction have a limitation, in that they cannot handle interactions between variables. In this section, LBC volume prediction methods are proposed by using two types of multivariate techniques that aim to resolve this shortcoming of univariate models and to improve prediction accuracy: manifold learning- and deep learning-based models.

3.3.1. Manifold learning-based model

Manifold learning-based models are a new approach for the port cargo volume prediction, as they combine manifold learning techniques and regression models. Manifold learning is a class of representation learning and dimensionality reduction that extracts a low-dimensional manifold to recover a high-dimensional space (Lin & Zha, 2008). It is generally divided into four types; global, local, linear, and nonlinear models (Ma & Fu, 2011). Our manifold learning-based approach for LBC volume prediction utilizes globally and locally linear embedding methods. As the globally linear embedding methods, principal components analysis (PCA) converts high-dimensional data to low-dimensional data using orthogonal transforms (Jolliffe, 2002) and multidimensional scaling (MDS) computes embeddings that attempt to preserve pairwise distance (Mead, 1992). As the locally linear embedding method, locally linear embedding (LLE) is a lower-dimensional projection method which preserves distances within local neighborhoods (Saul & Roweis, 2000). Then, the latent variables extracted from the manifold learning models are used to forecast the LBC volume through the use of two machine learning-based regression models, radial basis function (RBF) kernel-based support vector regression (SVR) (Lee et al., 2015) and random forests (RF). SVR is a state-of-the-art method for regression with nonlinear mapping capabilities of forecasting and RF is an ensemble learning method for regression by constructing a multitude of decision

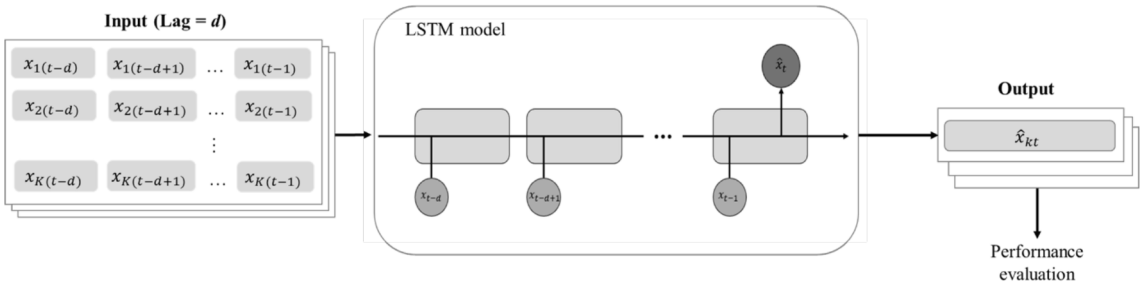


Fig. 7. LSTM for LBC volume prediction.

Table 1
Variable description of BL data.

Variable name	Description	Example
Ship name	Literal name of vessel	–
Call letter	Unique code of vessel	–
Number of arrivals	The number of vessel arrival	001
Date	Date of arrival in port	2017–01-01 01:20:00 AM
Facility name	Literal name of harbor facility	Dock
BL number	Unique codes for bill of lading	–
Ship company	Company name of vessel	–
Entry purpose	Classification of imports, exports, and transshipment	Import
Unloading port	Port name for unloading	Ulsan
Loading port	Port name for loading	Jiangyin
Unloading country	Country name for unloading	South Korea
Loading country	Country name for loading	China
HS code	Codes for categorizing items in accordance with international conventions	270,900
Item text	Name of goods for trading	Crude Oil
Weight ton	Measurement scale with weight	100.00
Volume ton	Measurement scale with volume	717.04
Unloading company	Company name for unloading	–
Consignor	Name of the subject who sends the items	–
Consignee	Name of the subject who receives the items	–
Notify party	Where to report the BL data	–

trees (Drucker et al., 1997; Ho, 1995). Of these, the manifold learning techniques coupled with the regression models have been considered, which are denoted PCA-SVR, PCA-RF, LLE-SVR, LLE-RF, MDS-SVR, and MDS-RF.

As shown in Fig. 6, input data $\boldsymbol{X} = [\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(K)}]$ is represented by horizontally concatenating cargo-related variables (i.e. variables of item and/or external data) $\boldsymbol{X}^{(k)} = [\boldsymbol{x}_1^{(k)}, \cdots, \boldsymbol{x}_{N-d+1}^{(k)}]^T, k = 1, \cdots, K$, which consist of time variables $\boldsymbol{x}_i^{(k)} = \left[\boldsymbol{x}_i^{(k)}, \cdots, \boldsymbol{x}_{i+d-1}^{(k)}\right]^T, i = 1, \cdots, N-d+1$; N is the sample size, K is the number of cargo-related variables, and d is the time lag. From the embedding techniques, latent variables $\boldsymbol{l}_z \in R^{N-d+1}, z = 1, \cdots, Z$ are extracted and used as new input variables for forecasting LBC volume by the regressors.

These models enable us to capture globally or locally linear relationships between $\boldsymbol{X}^{(k)}$ when learning representations. The non-linear relationship between $\boldsymbol{l}_z$ and the output from RBF kernel-based SVR and RF can be also identified. However, the manifold learning-based model considers time variables to be independent, and thus it cannot thoroughly reflect the effects and sequential characteristics of the time lag.

3.3.2. Deep learning-based model

Long Short-Term Memory (LSTM), one of the most widely-used recurrent neural networks (RNNs), was developed to prevent the gradient vanishing problem of the standard RNNs (Gers & Schmidhuber, 2001). Fig. 7 presents our data structure for the LSTM-based LBC volume prediction model. Input data as a 3-rank tensor consists of the batch size, time lag, and the number of cargo-related variables represented by *N, d,* and *K,* respectively. Many-to-one LSTM among various forms of LSTM was adopted for forecasting LBC volume because its objective is to forecast $\hat{x}_{kt}$ at time point t, right after the time lag of the given inputs (Fig. 7). Unlike manifold learning-based models, it reflects the time characteristic of the input data sequentially in model training.

4. Case study

The framework needs to be demonstrated for the following research questions: (a) what is the value of BL data in the proposed framework? (b) what are the effects of a correct item segmentation system on informative LBC volume analysis? (c) how can diverse data and the proposed prediction models affect LBC volume prediction accuracy? To do so, a case study is presented consisting of three parts: data description, experimental design, and results. Details of each part are given in the subsections below.

Table 2
Descriptive statistics of LBC volume in BL data.

Unit of volume: 1000 tons

	Crude oil		Oil product		Gas		Chemical		Plant & Animal oil	
year	Volume	Frequency	Volume	Frequency	Volume	Frequency	Volume	Frequency	Volume	Frequency
2007	270,304	401	102,243	1,971	18,745	310	78,737	6,848	93	57
2008	464,285	628	209,748	3,202	20,131	429	120,687	9,930	306	124
2009	459,299	620	220,015	3,320	22,078	392	123,437	9,180	209	118
2010	462,462	734	217,979	3,702	24,383	502	135,880	9,768	493	161
2011	526,109	838	275,895	4,535	23,725	530	140,277	9,877	541	159
2012	518,797	812	292,946	5,024	22,388	568	154,390	9,250	371	155
2013	492,959	802	297,427	5,015	23,254	572	145,307	8,851	342	104
2014	472,513	732	338,095	5,493	22,645	485	141,666	9,337	923	111
2015	487,742	683	321,524	5,205	29,182	489	131,360	10,282	1285	122
2016	506,786	719	340,622	6,253	43,641	609	131,753	11,547	1356	141
2017	513,841	705	358,018	7,065	42,899	580	138,793	12,599	1374	156

Table 3
Variable description of external data.

Category	Variable name
Balance of trade[1]	Export value index, Export volume index, Import value index, Import volume index, Net barter terms of trade index, Income trade condition
Exchange rates[2]	JPY/USD, USD/EUR, USD/GBP, KRW/USD, KRW/JPY, KRW/EUR
International reserves[3]	Gold price; Gold reserves, Foreign exchange reserves, Special drawing rights, IMF position
Balance of payments[4]	Commodity balance, Index of service, Primary income account, Transfer income account, Current account, Capital balance
International oil prices[5]	WTI, Dubai, Brent

Notes: [1]Balance of trade is the value related to transaction between countries. [2]Exchange rates are the relative values between two currencies (e.g. KRW/USD gives information how valuable the Korean currency compared to the US currency is.). [3]International reserves are external assets held by national central banks or a country's monetary authorities. [4]Balance of payments is the value calculated or balanced between profit and loss. [5]International oil price means the spot price of a barrel of crude oil. Detailed descriptions of each variable can be found in the ECOS.

Table 4
Example of LBC-ID.

Category	Subcategory	Keywords
Crude Oil	Crude Oil	Crude Oil, Condensate, Arabian, etc.
Oil Products	Gasoline	Gasoline, PYGAS, etc.
	Light Oil	Gas Oil, Diesel, MGO, LCO, etc.
	Asphalt	Asphalt, Bitumen
	Naphtha	Naphtha, Paraffin, Petroleum, etc.
	Fuel Oil	Fuel Oil, HSFO, LSFO, etc.
	Heavy Oil	Heavy Oil, Decant Oil, Bunker C
	Jet Oil	Jet
	Kerosene	Kerosene
	Base Oil	Base Oil, Lubricating, etc.
	Other.	Tudalen, Molten Sulphur, etc.
Gas	LPG Gas	LPG, Propylene, Butane, etc.
	LNG Gas	LNG, Gas Condensate, etc.
Chemical	Chemical	Paraxylene, Kokosol, Octene, etc.
Plant & Animal Oil	Plant & Animal Oil	Palm Oil, Coconut Oil, etc.

4.1. Data description

In this study, BL data from Ulsan port in Korea was used to analyze LBC volume; Ulsan port is the world's fourth-largest port in handling LBCs (Lee, 2015), and it is seeking to become an oil and gas hub in Northeast Asia. Hence, the data can represent the volume specific to LBCs. The LBC-related BL data consists of 172,954 samples from January 2007 to December 2017. Table 1 describes the variables extracted from the BL data for this study. Before applying proposed framework for Ulsan port, data preprocessing was conducted. In the preprocessing of volume data, 152 samples with null or error value of weight ton and volume ton were removed, and then a larger value between weight ton and 0.883 times volume ton was selected as LBC volume (Ulsan metropolitan city, 2019). After, in the case of item text data, the special characters and numbers were removed in item texts and capitalized letters. Afterward, item texts with numbers of characters between 2 and 30 were extracted. Finally, 172,802 LBC samples which consist of 15,249 unique item texts for item segmentation remained. Table 2 shows statistical analysis of BL data by category and year.

In addition, it is significant to consider the effects of external data which can influence LBC volume, since it can provide more informative prediction results. 26 monthly economic indicators, split into five categories over the period of 2007–2017 as the external data, were used. They were collected from economic statistics system (ECOS) in Korea (ECOS, 2018). Table 3 shows a description of the external data used in this study.

4.2. Experimental design

In this section, experimental design for proposed framework is introduced. For item segmentation part, LBC-ID partially shown in Table 4 was created for the ITS. The ITS was implemented for 15,249 preprocessed unique item texts. As mentioned before, LBC was generally classified into Crude Oil, Oil Products, Gas, Chemical, and Plant & Animal Oil at category level. Specifically, Oil Products could be divided into 10 subcategories: Gasoline, Light Oil, Asphalt, Naphtha, Fuel Oil, Heavy Oil, Jet Oil, Kerosene, Base Oil, and Other, which all involve different keywords.

After item segmentation and data transformation, BL data was converted into 132 monthly volume data. It was used for exploratory volume analysis and volume prediction. In detailed experimental setting of volume prediction, the data set was divided into training and test sets. The training set included 108 data records between 2007 and 2015, which is about 80% of the total data records. The test set consisted of 24 data records observed in the last two years of the study period (2016–2017). For this study, Oil Products and Light Oil were selected as target volume at category and subcategory level respectively.

Performance evaluation was conducted for LBC volume prediction. Two proposed prediction models were used to predict LBC volume aggregated from the ITS at both the category and subcategory levels, and our models with ARIMA, SARIMA, and Holts' Winters were compared. The performance of the prediction models was evaluated according to four prediction accuracy measures: root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and symmetric MAPE (SMAPE). Each model has user-defined parameters (i.e., hyperparameters) that can affect model performance. In this study, grid search was used to select an optimal value that minimizes RMSE and MAPE in the training set. The latent feature maps were explored at every five interval with the range of 10–30 and the search space of time lag was {3, 6, 12} for two proposed methods. Specifically, for training the deep learning-based model, experiments with diverse learning rates (i.e., {0.1, 0.01, 0.001}) and the number of epochs (i.e.,

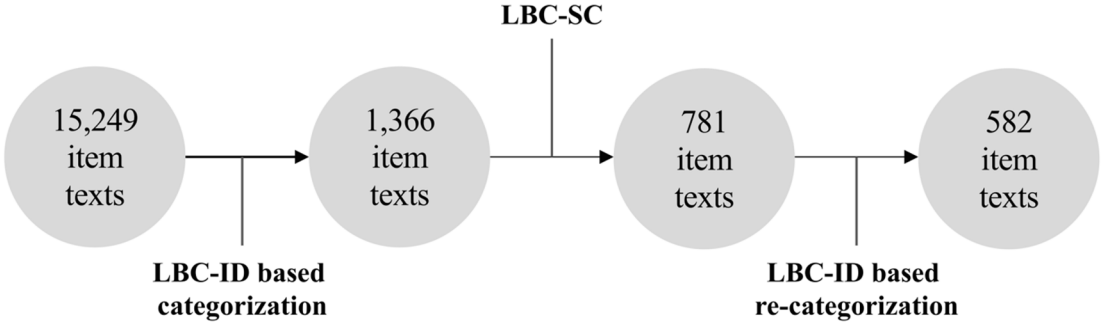


Fig. 8. Change of the number of unique item texts remaining after item segmentation.

Table 5
Example of item segmentation results.

HSCS			ITS			
Item text	HS code	Category	Item text (raw)	Item text (changed)	Subcategory	Category
Cuude Oil	270,900	Crude Oil	Cuude Oil	Crude Oil	Crude Oil	Crude Oil
Crude Oil	270,900	Crude Oil	Crude Oil	Crude Oil	Crude Oil	Crude Oil
Crude Oil	271,000	Oil Products	Crude Oil	Crude Oil	Crude Oil	Crude Oil
Jet	271,000	Oil Products	Jet	Jet	Jet Oil	Oil Products
Jet A 1	293,090	Chemical	Jet A 1	Jet A	**Jet Oil**	Oil Products
Asphalt	271,320	Oil Products	Asphalt	Asphalt	**Asphalt**	Oil Products
Asphlt	293,100	Chemical	Asphlt	Asphalt	**Asphalt**	Oil Products

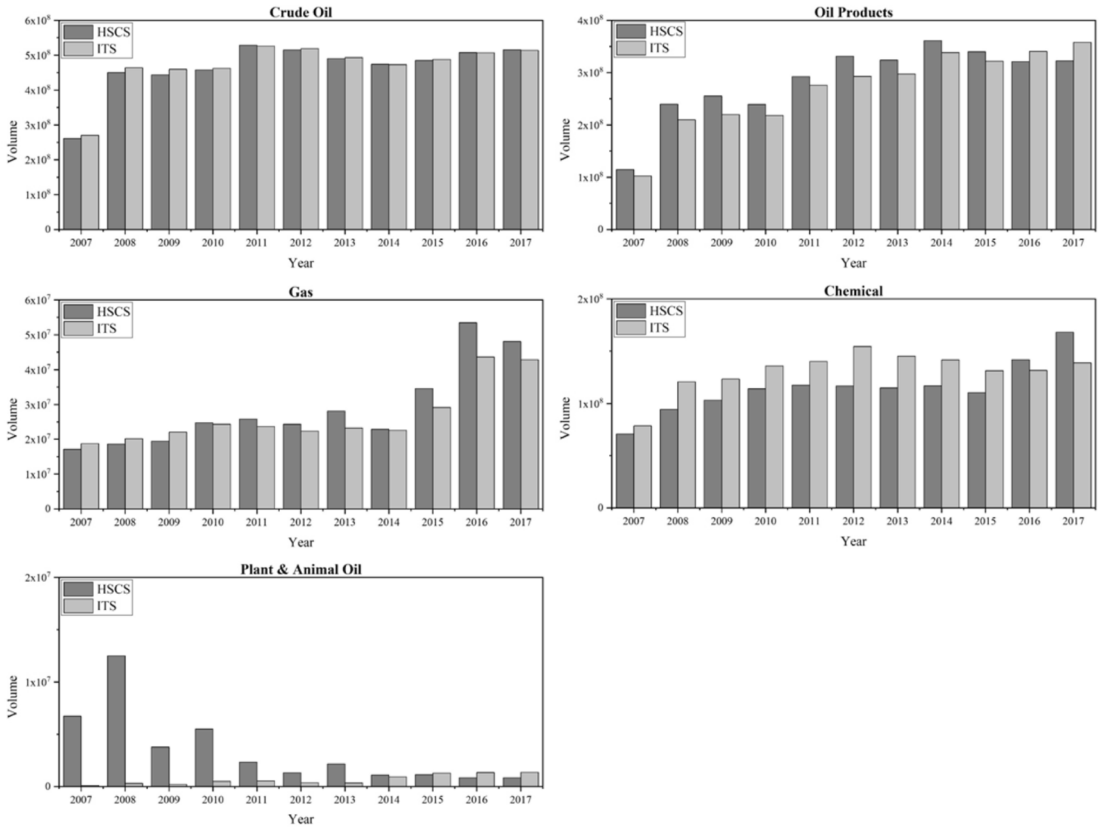


Fig. 9. Difference in LBC volume by category between ITS and HSCS.

{500, 1000, 2000}) were conducted. The optimal hyperparameters finally determined can be found in Appendix D.

Input data for LBC volume prediction was constructed by categorized volume data (Cat.), subcategorized volume data (Sub.cat.), and external data (Ext.), and input types were composed of diverse combinations of three data. To be specific, Cat. for Oil Products at the category level was defined as five categories of LBC, listed in the first column of Table 4. Sub.cat. for Oil Products and Light Oil at the subcategory level was defined as 10 subcategories of Oil Products in Table 4. In the case of the external data, only variables highly correlated with the target of LBC volume ($r \geq 0.5$) were used as shown in Table B1 Finally, for LBC volume prediction, experiments based on the following five combinations of input data were performed. Three combinations of input types (i.e., Input type 1 contains only Cat., Input type 2 – Cat. + Sub.cat., and Input type 3 – Cat. + Sub.cat. + Ext.) were used for Oil Products volume prediction at category-level. The other combinations (i.e., Input type 4 – Sub.cat. and Input type 5 –Sub.cat. + Ext.) were used for Light Oil volume prediction at subcategory-level. To see the effect of subcategory adjusted model for predicting the volume at the category level, Input types 2 and 3 consist of the lower level of volume data (Sub.cat.) as well as the same level volume (Cat.) compared with Input types 1 with only Cat. External data were added to Input types 3 and 5 in order to verify the effect of external data.

4.3. Results

4.3.1. Item segmentation

Fig. 8 shows the change in the number of unique item texts remaining after item segmentation. Approximately 96% of the 15,249 item texts was classified into the correct categories and subcategories by using the LBC-ID and the LBC-SC ($q \geq 10$). An actual example of item segmentation results can be found in Table 5, comparing the item segmentation results of the ITS with the HSCS. It can be seen that 'Cuude Oil' and 'Asphlt' which are typical typos in item texts (see the first and seventh row in Table 5), were correctly converted into 'Crude Oil' and 'Asphalt', respectively; 'Crude Oil', previously categorized into 'Oil Products' due to the incorrect HS code in the HSCS (see the third row in Table 5), was also classified into the correct category in the ITS.

The ITS is compared with the HSCS in terms of volume statistics estimation accuracy. Fig. 9 shows the difference in LBC volume in

Table 6
LBC volume change by year for subcategories of Oil Products.

	2013		2014		2015		2016		2017	
Rank	Sub-category	%	Sub-category	%	Sub-category	%	Sub-category	%	Sub-category	%
1	Lo	21.99	Lo	26.09	Lo	26.53	Lo	28.60	Lo	26.13
2	Jo	20.81	Jo	21.38	Jo	22.52	Fo	15.54	Fo	19.14
3	Na	15.81	Na	13.10	Gl	12.93	Jo	14.49	Jo	15.58
4	Gl	12.28	Gl	11.98	Na	12.31	Gl	13.80	Gl	11.75
5	E	10.28	E	8.67	Fo	8.89	Na	11.64	Na	10.14
6	Fo	8.78	Fo	8.56	E	6.63	E	5.90	E	7.09
7	As	3.62	As	3.36	As	3.89	As	3.99	As	4.10
8	Ks	3.49	Ks	3.28	Bo	3.16	Bo	2.79	Bo	3.20
9	Bo	2.81	Bo	3.08	Ks	3.14	Ks	2.67	Ks	2.70
10	Ho	0.13	Ho	0.51	Ho	0.00	Ho	0.56	Ho	0.17

Gl: Gasoline, Lo: Light oil, As: Asphalt, Na: Naphtha, Fo: Fuel oil, Ho: Heavy oil, Jo: Jet oil, Ks: Kerosene, Bo: Base oil, E.: etc.

Table 7
LBC volume change of Light Oil by purpose of shipment and area for three years.

Unit: 1000 tons

	Import			Export			Transshipment		
	2015	2016	2017	2015	2016	2017	2015	2016	2017
Far East	836	1,347	5,407	5,697	11,377	18,173	116	418	294
South America	–	–	–	–	1,920	855	–	–	–
Oceania	–	650	–	7,331	6,954	13,522	–	–	–
South East Asia	–	115	–	50,394	37,285	34,138	–	–	–
North America	–	–	359	–	764	180	–	–	–
Western Asia	40	–	–	330	8,886	16,880	–	–	–
Africa	–	–	–	2,572	9,272	2,262	–	–	–
Europe	–	–	–	6,406	3,673	–	–	–	–
Japan	922	1,818	1,336	2,542	4,128	2,918	381	–	–
Middle East	–	–	512	1,486	3,254	–	–	–	–
Central America	–	–	–	–	–	–	–	–	–
Other	–	–	747	233	253	600	–	–	–

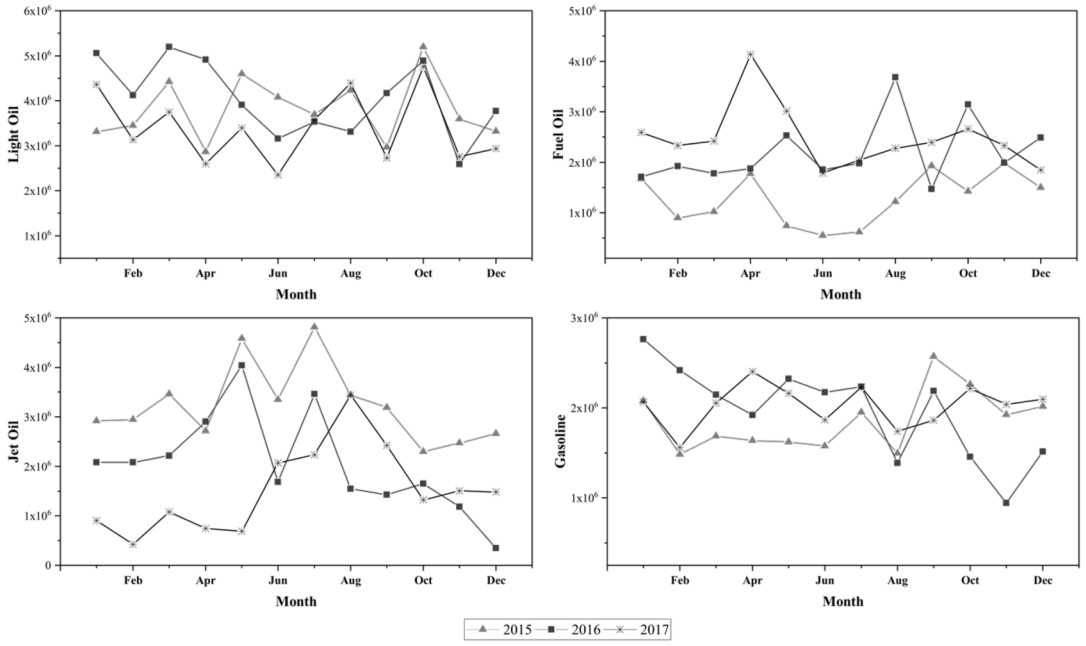
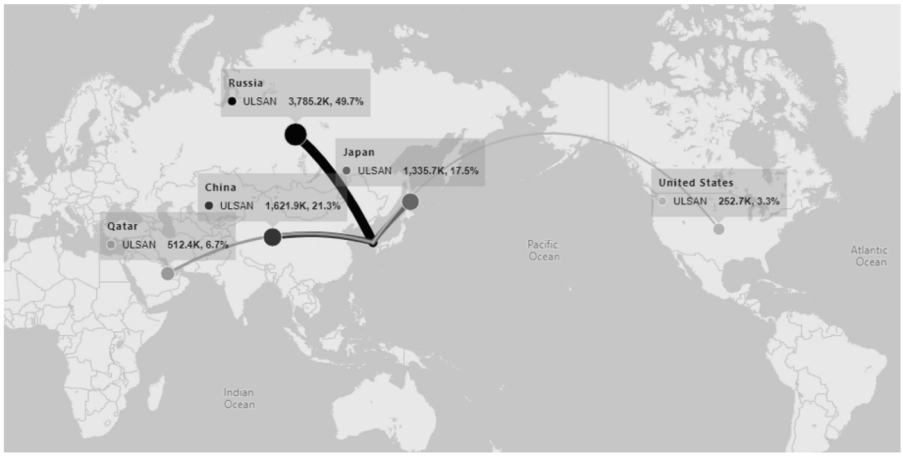


Fig. 10. Monthly trends for LBC volume of top-4 subcategories of Oil Products by year.

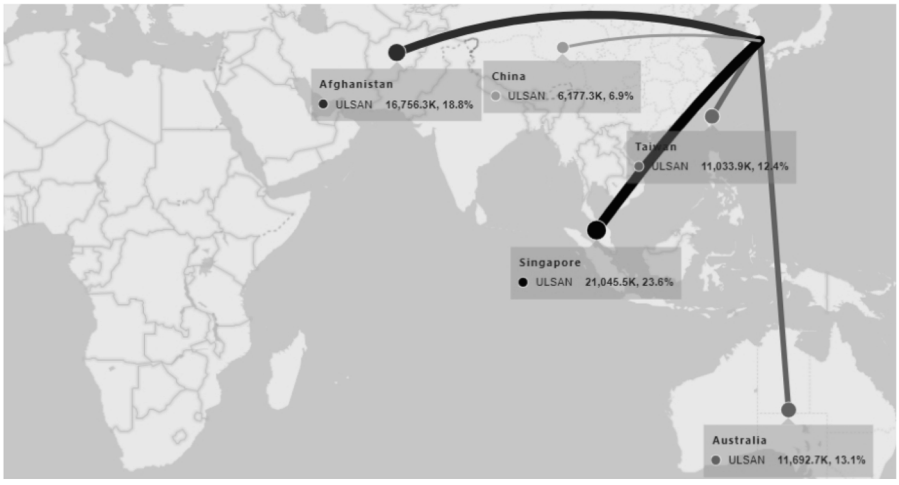
categories between the ITS and the HSCS. For example, there is a significant gap in Plant & Animal Oil, which might be the result of errors in the existing system for Crude Oil and/or Oil Products. As a result, it can be demonstrated to the usefulness of the ITS, and the importance of correct LBC volume aggregation based on the ITS.

4.3.2. *Exploratory volume analysis*

Exploratory volume analysis for LBC volume in terms of statistical view was firstly conducted to capture its trends. Table 6 shows the percentage of LBC volume for subcategories of Oil Products from 2013 to 2017. As an example, Light Oil makes up the highest proportion over the



(a) Light Oil volume from top-5 countries to Ulsan in 2017 for import



(b) Light Oil volume from Ulsan to top-5 countries in 2017 for export

Fig. 11. Flow map graphs for Light Oil volume by country in 2017 (a) Light Oil volume from top-5 countries to Ulsan in 2017 for import (b) Light Oil volume from Ulsan to top-5 countries in 2017 for export.

five years. In 2015, there are several changes in the rankings. Gasoline outpaced Naphtha, while Fuel Oil climbed one place with a marginal increase. Compared to the small increase of Fuel Oil in the previous year, it increased sharply to surpass Jet Oil in 2016.

Table 7 shows the changes in the volume of Light Oil for three years by purpose of shipment and area. In the case of Light Oil, the export volume is much higher than the import volume, and it is mainly imported from the Far East, Oceania, and Japan, and exported to most areas. In the last three years, transshipment volume has only occurred in the Far East and Japan.

Fig. 10 identifies the monthly volume trends of top-4 subcategories in Oil Products in terms of timeline view. For Light Oil, the volume for all three years declined in February and increased in March, while gasoline decreased in August and increased in September.

In port planning and management, it is important to identify major trading countries. As an example, Fig. 11 shows the top-5 countries which trade Light Oil in terms of imports and exports through Ulsan port in by 2017 based on geographic view. Regarding imports of Light Oil, Russia is the dominant country which accounted for approximately 50% of Light Oil imports into Korea by volume. Meanwhile, most of the volume of Light Oil for export goes to the top-5 countries with their volume accounting for about 75% of the total volume, even though no particular country is dominant. Interestingly, China is also considered as a major country for Light Oil trade with Korea since it records top-tier ranking for both imports and exports. It is noted that these analyses are available for any items as well as Oil Products and Light Oil.

4.3.3. *Volume prediction*

Figs. 12 and 13 show LBC volume prediction results for Oil Products (category level) and Light Oil (subcategory level), respectively. In the case of Oil Products, five categories, 10 subcategories of Oil Products, and 14 variables of external data were used as factors; for Light Oil, there were 10 subcategories of Oil Products and eight variables of external data, which are presented in Table 3 and B1. In Fig. 12, LSTM is shown to outperform ARIMA, SARIMA, and Holts' Winters in all prediction accuracy measures; in particular, the performance of LSTM in the

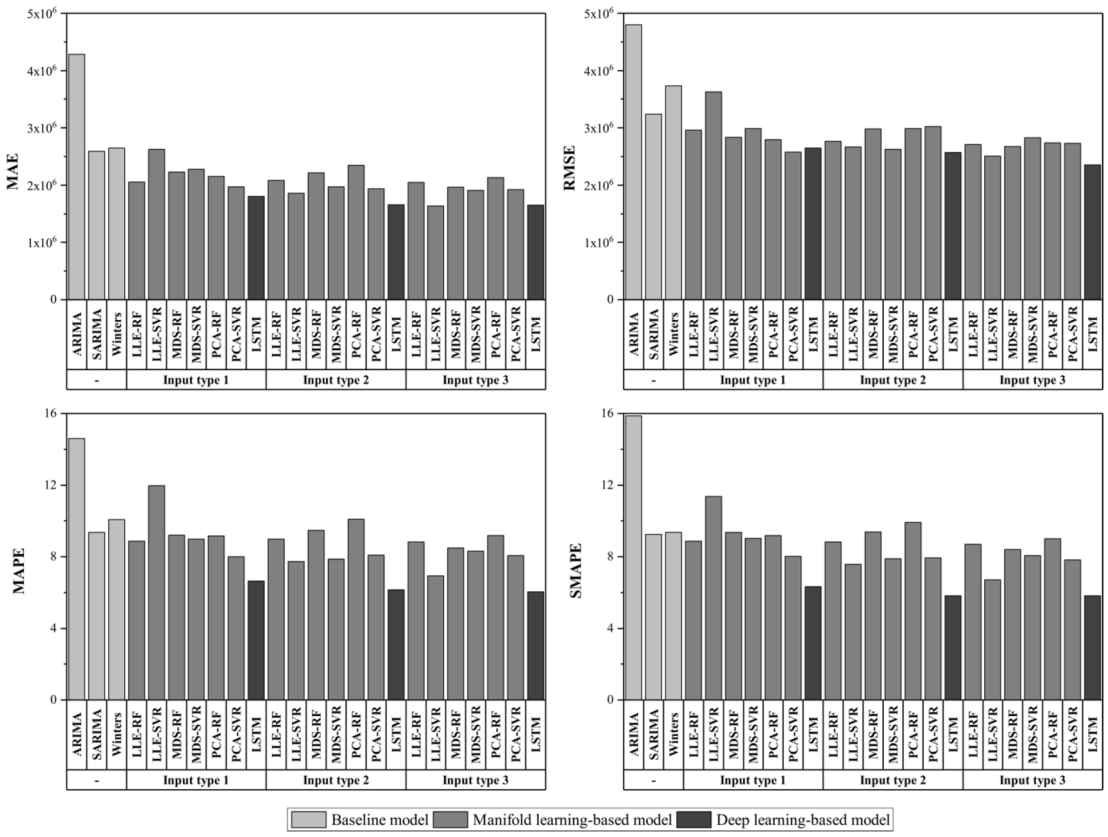


Fig. 12. Prediction results for Oil Products volume (category level).

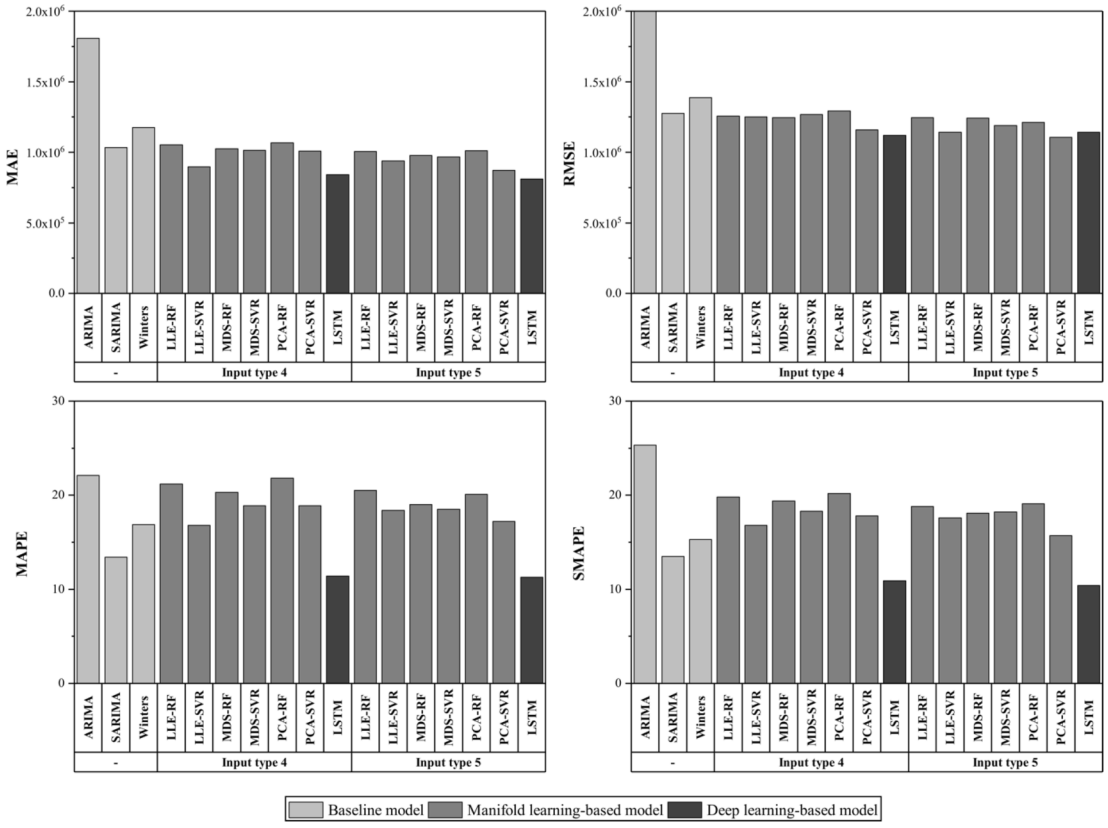


Fig. 13. Prediction results for Light Oil volume (subcategory level).

experiment using input type 3 is the best.

Fig. 13 shows that manifold learning-based models are partially better than the traditional models, but LSTM outperforms the traditional ones in all aspects. In particular, the LSTM that used input type 5 reflecting subcategories and external data shows better performance than input type 4 reflecting only subcategories in all prediction measures. From the prediction results of Oil Products and Light Oil, LSTM, a deep learning-based model, displays superior performance to the other nine prediction models. Noticeably, it was demonstrated that the experiments for LSTM using the subcategories previously generated from the ITS show a greater improvement in predictive power than models that use only one item, or items at the category level only. Overall LBC volume prediction that includes the external data shows better predictive power than models without the external data. The specific figures of performance for each model in the experiments are represented in Table E1 and E2.

5. Discussion

The numerical results of our proposed framework can provide several methodological implications. Accurate item segmentation for LBCs with ITS (Fig. 8) leads to improving the volume aggregation accuracy. As shown in Fig. 9, in the case of Plant & Animal oil in 2008, about 12 million tons of error caused by miswriting input were corrected, and then its desired trend changes could be captured; Plant & Animal oil was decreasing before conducting the ITS, but it showed a trend of gradually increasing after the ITS. To sum up, the ITS is an innovative system that can improve the quality and reliability in aggregating LBC volume statistics by conducting item categorization systematically and efficiently as valid categories and subcategories. Furthermore, port cargo volume generally depends on time, cargo type, and other cargoes; our proposed methods, which are multivariate time series models with various input data types, can effectively model these relationships. In particular, the effects of external data on volume prediction were explored. The accuracy of volume prediction based on input data including external data increased by 34% and 18% in average at category and subcategory levels over baseline models, respectively. It implies that the proposed model can reflect the fluctuation of LBC volume which is heavily influenced by economic indicators due to cross-border relations and circumstances.

Meanwhile, this study also provides several interesting managerial insights. First, this decision support system for LBC volume analysis has various potential users such as port-related agencies, port officials and experts, and maritime logistics companies. It can provide information that enable such stakeholders to respond to changes in the international situation more quickly by identifying the logistics movements between countries or geographic areas and helping build precise knowledge about the potential strategic cargoes by analyzing the LBC volume down to the subcategory level. For instance, the Naphtha transaction ratio at Ulsan Port is declining from 15.81% in 2013 to 10.14% in 2017 (Table 6). During that period, international naphtha prices declined, leading to a decline in operating profits for domestic refineries that mainly export naphtha. Because companies sold Ethylene instead of Naphtha for increasing profits, Naphtha volume declined (Jeon, 2018). Thus, considering the situation, oil refining companies can maintain the price competitiveness by controlling the supply of LBC cargoes such as Naphtha and Ethylene. Furthermore, the flows of LBC volume between countries and ports are identified in Fig. 11 based on the geographic visualization results of exploratory volume analysis.

From this, ports can provide a reduction in port-dues based on cross-border relations and seaways. It may help to expand the usage rates of ports by attracting new customers and transshipment cargo and continuously increases the activation of new docks (Lee, 2019). Our framework also can help oil refining or petrochemical companies find new opportunities relevant to market expansion through regional cargo volume analysis. Such companies need information on LBC volume for their production and import/export plans, and they may also decide to purchase sites for refining facilities or to invest in companies that are associated with a specific item of cargo based on the information provided by our models. In addition, since many cargo-specific facilities are often needed to (un)load various types of cargo for each port (e.g., the Ulsan port is specialized in handling LBCs), this framework can make a significant contribution to the decision-making processes of governments and public institutions charged with of building infrastructure and devising operating strategies.

6. Conclusion

As international trade is more active now than ever before, the smartization of ports is becoming increasingly important. In this study, the novel framework for LBC volume analysis was proposed, which can accelerate the transformation into smart ports and provide useful information for LBC-related port planning. This study breaks the existing volume analysis paradigm that has simply predicted the trends of future cargo volume and gives a direction for new research about port cargo volume analysis.

Although our automated framework is useful in analyzing LBC volume, there are some limitations in our study that should be considered in future work. Firstly, the LBC-ID has a reliability issue since it was manually constructed by considering several criteria based on information related to the petroleum industry. Secondly, the remaining unique item texts after item segmentation still require further processing; it is necessary to keep refining the LBC-ID by adding the keywords in consultation with domain experts in oil refining and petrochemical processes. Lastly, the usefulness of the generated subcategories and external data was demonstrated in the predictive analysis through various experiments that changed the input types, but it is still very difficult to identify the direct cause of the change in the volume of each item.

Several further research directions are as follows. First, this study can be applied for any type of cargoes, not only LBCs. Provided that a proper item dictionary specific to each cargo is in place, it can be extended to air logistics (Zou et al., 2013). Second, by presenting more specific subcategories, this study provides researchers with a guideline that can extend the scope of cargo volume analysis in depth beyond the total or category levels of LBC volume. Finally, the framework of this study can be a reference for research related to the development of item text-based item segmentation algorithms specialized to work with BL data, and it may lead to new research that clarifies the HS code or re-establishes the HSCS by recognizing the current ambiguities in the HS code.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Research Foundation (NRF) grant funded by the Korea Government (MSIT) under Grant No. 2020R1C1C1011063. This work was also supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the MIST (No. 2020-0-01336, Artificial Intelligence graduate school support (UNIST)) and by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5A2A03041137). Also, This work was supported by the 2018 Research Fund (1.180060.01) and the 2020 Research Fund (1.200098.01) of UNIST. This work was a part of the project titled 'Smart Maritime Accelerator Center,' funded by the Ulsan Port Authority (Grant No. 2.190533.01), Korea.

Appendix A

See Table A1.

Table A1
LBC-ID for oil products.

Subcategory	Keywords
Gasoline	Gasoline, PYGAS, Alkylate, Motor Spirit, Petrol, Pyrolysis Gasoline, Mogas, Avgas, Aviation Spirit, Carbob, Reformate
Light Oil	Gas Oil, Diesel, Light Cycle Oil, MGO, LCO
Asphalt	Asphalt, Bitumen
Naphtha	Naphtha, Paraffin, Petroleum, Aromatic, Mineral Spirit, Benzine, Hexane, Ligroin, White Oil, White Gas, Isopar, VMP
Fuel Oil	Fuel Oil, HSFO, LSFO, SRFO, MF Cst, Neutral Oil, Engine Oil, Stadis
Heavy Oil	Heavy Oil, Decant Oil, Bunker C
Jet Oil	Jet
Kerosene	Kerosene
Base Oil	Base Oil, Lubric, Lube, Ultras, Yubase, Kixxlubo, Sellmvin, Hydro Cracker Bottoms, UCO, Oil Base, Gear Oil, Base Cst, Lubrizol
Etc.	Tudalen, Molten Sulphur, Mixture, Spindle, Refined Oil, Roll Oil, Annex Oil, Process Oil, Cutting Oil, Machine Oil, Transformer Oil, Unconveried Oil, Insulation Oil, Oil Extender, Norman, SKN

Appendix B

Table B1 indicates the highly correlated external variables with each target variable. The variables were selected by Pearson coefficient correlation that can measure the linear dependency between two variables (Lee, Choi, & Jun, 2021).

Table B1
Highly correlated external data variables with Oil Products and Light Oil volume.

Correlation range	For Oil Products	For Light Oil
$0.5 \leq \lvert r \rvert$	Export value index (0.72); Export volume index (0.85); Import volume index (0.78); Income trade condition (0.76); USD/GBP (-0.73); Foreign exchange reserves (0.80); Import value index (0.51); USD/EUR (-0.59); Gold price (0.59); Gold reserves (0.80); Special drawing rights (0.62); IMF position (0.70); Current account (0.61); Commodity balance (0.63)	Export volume index (0.62) Import volume index (0.68) Income trade condition (0.69) USD/EUR (-0.64) USD/GBP (-0.64) Gold reserves (0.61) Foreign exchange reserves (0.63) Commodity balance (0.57)

Appendix C

See Table C1.

Table C1
Correlation of the segmented items.

	Co	Op	Ga	Ch	Pa	Gl	Lo	As	Na	Fo	Ho	Jo	Ks	Bo	Etc.
Co	1.00														
Op	0.57	1.00													
Ga	0.49	0.11	1.00												
Ch	0.71	0.67	0.29	1.00											
Pa	0.52	0.18	0.45	0.32	1.00										
Gl	0.92	0.72	0.45	0.92	0.45	1.00									
Lo	0.78	0.16	0.49	0.48	0.55	0.65	1.00								
As	0.89	0.55	0.50	0.58	0.54	0.80	0.67	1.00							
Na	0.52	0.44	0.01	0.42	0.09	0.50	0.08	0.44	1.00						
Fo	0.62	0.28	0.60	0.34	0.38	0.52	0.52	0.59	0.06	1.00					
Ho	0.83	0.58	0.46	0.70	0.46	0.83	0.66	0.75	0.32	0.50	1.00				
Jo	-0.10	-0.13	0.07	-0.23	0.04	-0.17	-0.01	-0.10	-0.12	0.01	-0.20	1.00			
Ks	0.33	0.42	-0.16	0.40	0.12	0.40	0.06	0.21	0.30	-0.22	0.20	-0.21	1.00		
Bo	0.40	0.20	0.09	0.29	0.03	0.36	0.32	0.29	0.31	0.25	0.32	-0.02	-0.19	1.00	
Etc.	0.64	0.39	0.36	0.36	0.46	0.55	0.44	0.67	0.40	0.39	0.48	-0.01	0.20	0.16	1.00

Abbreviations: Co: Crude oil, Op: Oil product, Ga: Gas, Ch: Chemical, Pa: Plant & Animal oil, Gl: Gasoline, Lo: Light oil, As: Asphalt, Na: Naphtha, Fo: Fuel oil, Ho: Heavy oil, Jo: Jet oil, Ks: Kerosene, Bo: Base oil, E: Etc.

Appendix D

See Table D1.

Table D1
Hyperparameters used in Oil Products and Light Oil prediction.

Predicted item	Models	*p*	*d*	*q*	Lag	Layer	Hidden layer	Learning rate	Epochs
Oil Products	ARIMA	1	0	1	–	–	–	–	–
	SARIMA	1	0	1	–	–	–	–	–
	Winters	–	–	–	12	–	–	–	–
	PCA	–	–	–	12	–	15	–	–
	LLE	–	–	–	12	–	15	–	–
	MDS	–	–	–	12	–	15	–	–
	LSTM	–	–	–	6	1	30	0.1	1000
Light Oil	ARIMA	1	0	1	–	–	–	–	–
	SARIMA	1	0	1	–	–	–	–	–
	Winters	–	–	–	12	–	–	–	–
	PCA	–	–	–	12	–	20	–	–
	LLE	–	–	–	12	–	20	–	–
	MDS	–	–	–	12	–	20	–	–
	LSTM	–	–	–	12	1	30	0.01	100

The optimal hyperparameters were selected based on Akaike information criterion for ARIMA, SARIMA, and Winters; RMSE and MAPE for PCA, LLE, MDS, and LSTM.

Appendix E

See Table E1 and E2.

Table E1
Prediction results for Oil Products volume.

Input type	Model type	Models	MAE	RMSE	MAPE	SMAPE
	Baseline model	ARIMA	4,285.4	4,797.3	14.6	15.9
		SARIMA	2,590.7	3,235.8	9.4	9.2
		Winters	2,647.4	3,729.9	10.1	9.4
Input type 1 (Cat.)	Manifold learning-based model	LLE-RF	2,056.2	2,958.9	8.9	8.9
		LLE-SVR	2,622.8	3,627.8	12.0	11.4
		MDS-RF	2,228.4	2,832.0	9.2	9.4
		MDS-SVR	2,275.3	2,989.1	9.0	9.0
		PCA-RF	2,155.2	2,793.3	9.2	9.2
		PCA-SVR	1,970.2	**2,575.6**	8.0	8.0
	Deep learning-based model	LSTM	**1,802.0**	2,644.4	**6.6**	**6.3**
Input type 2 (Cat. +Sub.cat.)	Manifold learning-based model	LLE-RF	2,080.9	2,766.8	9.0	8.8
		LLE-SVR	1,856.0	2,669.1	7.7	7.6
		MDS-RF	2,211.8	2,982.7	9.5	9.4
		MDS-SVR	1,968.7	2,623.8	7.9	7.9
		PCA-RF	2,343.8	2,988.1	10.1	9.9
		PCA-SVR	1,934.8	3,021.7	8.1	7.9
	Deep learning-based model	**LSTM**	**1,656.8**	**2,570.7**	**6.2**	**5.8**
Input type 3 (Cat. +Sub.cat. +Ext.)	Manifold learning-based model	LLE-RF	2,046.4	2,710.9	8.8	8.7
		LLE-SVR	1,739.8	2,508.1	6.9	6.7
		MDS-RF	1,965.8	2,671.6	8.5	8.4
		MDS-SVR	1,909.0	2,823.9	8.3	8.1
		PCA-RF	2,131.2	2,734.2	9.2	9.0
		PCA-SVR	1,922.5	2,728.3	8.1	7.8
	Deep learning-based model	**LSTM**	**1,652.1**	**2,351.1**	**6.0**	**5.8**

The unit of MAE and RMSE is 1000.

Table E2
Prediction results for Light Oil volume.

Input type	Model type	Models	MAE	RMSE	MAPE	SMAPE
	Baseline model	ARIMA	1,807.4	2,009.6	22.1	25.3
		SARIMA	1,034.7	1,275.3	13.4	13.5
		Winters	1,176.6	1,386.5	16.9	15.3
Input type 4 (Sub.cat.)	Manifold learning-based model	LLE-RF	1,052.4	1,255.9	21.2	19.8
		LLE-SVR	897.2	1,250.3	16.8	16.8
		MDS-RF	1,024.5	1,244.3	20.3	19.4
		MDS-SVR	1,013.1	1,266.3	18.9	18.3
		PCA-RF	1,066.2	1,293.7	21.8	20.2
		PCA-SVR	1,009.2	1,159.1	18.9	17.8
	Deep learning-based model	**LSTM**	**842.2**	**1,120.2**	**11.4**	**10.9**
Input type 5 (Sub.cat. +Ext.)	Manifold learning-based model	LLE-RF	1,006.4	1,245.6	20.5	18.8
		LLE-SVR	939.9	1,143.0	18.4	17.6
		MDS-RF	978.3	1,242.0	19.0	18.1
		MDS-SVR	967.6	1,188.6	18.5	18.2
		PCA-RF	1,011.7	1,211.0	20.1	19.1
		PCA-SVR	872.3	**1,106.7**	17.2	15.7
	Deep learning-based model	**LSTM**	**810.1**	1,140.9	**11.3**	**10.4**

The unit of MAE and RMSE is 1000.

Appendix F

Performance evaluation was conducted by comparing the effect of the ITS to that of the HSCS in terms of volume prediction. The prediction models were implemented by using two training sets: one from the HSCS (i.e. HSCS training set) and the other from the ITS (i.e. ITS training set). Then, each trained model was applied to the same test set from the ITS (i.e. ITS test set), which considered to be the correct results that the models should predict since it is confirmed that the ITS is an accurate system since it corrected the errors of the HSCS in Section 4.3.1 and 4.3.2. At this time, the trained models, which showed similar performance during training were utilized for a fair performance comparison. Table F1 presents the prediction results of LSTM for Oil Products volume when using the HSCS-based and the ITS-based training sets respectively (i.e., HSCS-LSTM and ITS-LSTM). ITS-LSTM outperforms HSCS-LSTM in all four prediction measures. This shows that the ITS allows more accurate LBC volume prediction than the HSCS by correcting the errors derived from the HSCS.

Table F1
Performance comparison of HSCS-LSTM and ITS-LSTM.

Model	MAE	RMSE	MAPE	SMAPE
HSCS-LSTM	4,620.5	4,838.4	14.6	15.7
ITS-LSTM	**2,559.2**	**2,934.6**	**8.9**	**9.0**

The unit of MAE and RMSE is 1000.

References

Adland, R., Jia, H., & Strandenes, S. P. (2017). Are AIS-based trade volume estimates reliable? The case of crude oil exports. *Maritime Policy and Management., 44*(5), 657–665. https://doi.org/10.1080/03088839.2017.1309470

Drucker, H., Surges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*.

ECOS (2018). ECOS Meta DB. Economic Statistics System of the Bank of Korea. Available at: https://ecos.bok.or.kr/jsp/use/metaword/MetaDataWordListPopUp.jsp.

Fagerholt, K., Kim, K.-H., Lee, C.-Y., Meng, Q., & Qi, X. (2017). Editorial of special issue on ocean transportation logistics: Making global supply chain effective. *Flexible Services and Manufacturing Journal, 29*(3-4), 309–311. https://doi.org/10.1007/s10696-017-9293-7

Fjærtoft, D. B. (2015). Modeling Russian regional economic ripple effects of the oil and gas industry: Case study of the republic of Komi. *Regional Research of Russia, 5*(2), 109–121. https://doi.org/10.1134/S2079970515020033

Gers, F. A., & Schmidhuber, E. (2001). LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks. doi, 12*(6), 1333–1340.

Guszczak, B., & Mencarelli, R. (2020). *IT Solutions Supporting Information Exchange in Intermodal Transport.*. https://doi.org/10.1007/978-3-030-24355-5_9

Ho, T. K. (1995). Random decision forests. *Proceedings of the International Conference on Document Analysis and Recognition*. https://doi.org/10.1109/ICDAR.1995.598994

Jai Sankar, T., Vijayalakshmi, C., & Poovaraaghavan, J. (2016). Design of forecasting model for export of liquid bulk. *International Journal of Pure and Applied Mathematics*.

Jeon, J. S. (2018). An Analysis of the Oil Industry's Management Performance and Future Prospects Report in 2017 (Vol. 154).

Jolliffe, I. T. (2002). Principal component analysis. *Encyclopedia of statistics in behavioral science* (second ed.) 10.2307/1270093.

Kim, H.-T., & Ko, B.-W. (2007). A Study for Increasing Value-Added by Developing Korean Chemical Ports into East-Asian Hub Ports.

Kim, J.-E., Oh, J.-H., & Woo, S.-H. (2018). An introduction of new time series forecasting model for oil cargo volume. *Journal of Korea Port Economic Association, 34*(1), 81–98.

Kim, J.-E., & Woo, S.-H. (2017). Forecasting oil freight volume of the port of Ulsan. *Korea International Commerce Review, 32*, 91–108.

Kim, J.-H. (2008). The forecast of the cargo tranportation for the North port in Busan, using time series models. *Journal of Korea Port Economic Association, 24*(2), 1–17.

Lee, D. (2019). A Study on the Analysis of the Effects of Port Use Fee Reduction System for Liquid Cargo of Ulsan Port.

Lee, H. S. (2015). A Study on the Development Strategy of Ulsan Port Authority using SWOT/AHP Analysis [Korea Maritime and Ocean University]. http://kmou.dcollection.net/jsp/common/DcLoOrgPer.jsp?sItemId=000002174452.

Lee, J. (2020). The Effects of Conflict on South and North Korean Economy : Evidence from Stock Market and Foreign Trade.

Lee, J., Chang, K., Jun, C. H., Cho, R. K., Chung, H., & Lee, H. (2015). Kernel-based calibration methods combined with multivariate feature selection to improve accuracy of near-infrared spectroscopic analysis. *Chemometrics and Intelligent Laboratory Systems, 147*, 139–146.

Lee, J., Choi, I. Y., & Jun, C. H. (2021). An efficient multivariate feature ranking method for gene selection in high-dimensional microarray data. *Expert Systems With Applications, 166*, 113971.

Lee, P. T.-W., & Lee, T.-C. (2016). New Concepts in the Economies of Flow, Connection, and Fusion Technology in Maritime Logistics. *Dynamic Shipping and Port Development in the Globalized Economy*. https://doi.org/10.1057/9781137514295_9.

Lee, S., Song, J., Park, S., & Sohn, B. (2014). A study on the comparative analysis of port competitiveness using AHP. *KMI International Journal of Maritime Affairs and Fisheries.*

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*.

Lin, T., & Zha, H. (2008). Riemannian manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(5), 796–809. https://doi.org/10.1109/TPAMI.2007.70735

Ma, Y., & Fu, Y. (2011). *Manifold Learning Theory and Applications* (1st ed.). CRC Press. https://doi.org/https://doi.org/10.1201/b11431.

Mead, A. (1992). Review of the development of multidimensional scaling methods. *The Statistician, 41*(1), 27. https://doi.org/10.2307/2348634

Merk, O., Busquet, B., & Aronietis, R. (2015). The impact of mega-ships: case-specific policy analysis. *OECD/International Transport Forum*. https://doi.org/10.1093/innovait/inr040

Peterson, J. L. (1980). Computer programs for detecting and correcting spelling errors. *Communications of the ACM, 23*(12), 676–687. https://doi.org/10.1145/359038.359041

Saul, L., & Roweis, S. (2000). An introduction to locally linear embedding. Unpublished. Available at: http://Www. Cs. Toronto.https://doi.org/10.1016/j.jtbi.2004.07.023.

Spyridoula, K. C. (2019). TRANSSHIMENT HUBS : THE CASE OF MEDITERRANEAN PORTS AND THE PROSPECTS. October.

Ulsan metropolitian city. (2019). The statistical yearbook of Ulsan.

Wang, D., Li, D., Gong, Y.u., Wang, R., Wang, J., & Huang, X. (2019). Development situation and future demand for the ports along the Northern Sea Route. *Research in*

Transportation Business and Management, 33, 100465. https://doi.org/10.1016/j.rtbm.2020.100465

C.H. Yu Exploratory Data Analysis 1977 Methods In Oxford Bibliographies 10.1093/OBO/9780199828340.

Zhang, Z., & Xing, W. (2018). Overseas Oil Cooperation between China and India Based on Crude Oil Trade Flow Analysis. IOP Conference Series: Earth and Environmental Science. https://doi.org/10.1088/1755-1315/153/3/032046.

Zhou, C., Li, H., Liu, W., Stephen, A., Lee, L. H., & Chew, E. P. (2019). Challenges and opportunities in integration of simulation and optimization in maritime logistics. *Proceedings-Winter Simulation Conference*. https://doi.org/10.1109/WSC.2018.8632202

Zou, L.i., Yu, C., & Dresner, M. (2013). The application of inventory transshipment modeling to air cargo revenue management. *Transportation Research Part E: Logistics and Transportation Review., 57*, 27–44. https://doi.org/10.1016/j.tre.2013.01.004