A data-driven optimization of large-scale dry port locations using the hybrid approach of data mining and complex network theory

Abstract

The paper proposes a two-stage approach that combines data mining and complex network theory to optimize the locations and service areas of dry ports in a large-scale inland transportation system. In the first stage, candidate locations of dry ports are weighted based on their eigenvector centrality in the complex network of association rules mined from a large amount of international transaction data. In the second phrase, dry port locations and their service areas are optimized using the gravity-based community structure. The method is validated in a real case study which optimizes a large-scale dry port network in Mainland China in the context of the Belt and Road Initiatives (BRI). As a result, optimal dry port locations include key transportation hubs that closely reflect the real BRI development plan, hence, the proposed approach is validated.

Keywords: Transportation, Data mining; Large scale optimization; Dry ports; Complex network theory

1. Introduction

With the rapid development of globalization and international trade, intercontinental freight transport has experienced a fast-paced growth rate of 9.3% per year, from just under 85 million twenty-foot equivalent units (TEUs) in 1990 to about 651 million TEUs in 2013 (Lee and Song, 2017). Nevertheless, as container flows continue to rise steeply, many seaports have been confronted with the problem of severe congestion in terminals and bottlenecks in the inland transportation system (Chang et al., 2015). Under such circumstances, dry ports have been increasingly implemented as an effective logistics solution to sustain seaport competitiveness and improve the efficiency of the freight transportation chain as a whole (Qiu and Lee, 2019).

By definition, dry ports are inland intermodal terminals connected directly to one or several seaports by high-capacity transport modes, preferably railways, where shippers and carriers can drop off and/or pick up their containers directly as if going to seaports (Crainic et al., 2015). In general, dry ports provide almost all services offered at a seaport, such as customs clearance, storages, maintenance and repair of empty containers, tax payments, and other value-added logistics activities. By transferring these services to the hinterland, dry ports can help ease many pressures and constraints faced by seaports, such as alleviating congestion at terminals and surrounding areas, increasing berth throughputs, improving inland accessibility, as well as offering better services to shippers and transport operators (Roso and Lumsden, 2010; Wang and Meng, 2019).

Operating as a consolidation point and logistics hub in the broader transport network, the success of a dry port is critically dependent on its location advantage (Lättilä et al., 2015). A well-selected location

can help dry ports attract adequate freight volumes from inland shippers, attaining economies of scale with full train services to seaports (Roso et al., 2009). Conversely, poorly planned dry ports can result in overcapacity, facility redundancy, a low efficiency and utilization rate, and threatening returns on investment. More importantly, once a dry port is built, it is almost impossible to relocate because of the heavy capital investment involved and the location-bound and sunk cost nature, as Chang et al. (2015) explain. Therefore, it is imperative to optimize the location and coverage area of dry ports at an early stage of their development.

Recent research has made good progress in applying most of the traditional modelling approaches, such as multi-criteria decision making (MCDM) and mixed-integer programming (MIP) approach, from facility location theory to dry port developments (Chang et al., 2015; Witte et al., 2019). However, most of exiting research addresses the dry port location problem on a small scale, while the large-scale optimization of dry port locations is still understudied.

With the fast growing availability of big data and recent advances in machine learning methodologies, both academics and practitioners have been increasingly paying attention to the development of the data-driven supply chain (SC) capabilities for better operational and financial performance (Yu, Chavez, et al., 2018). As a result, data-driven applications have been used to address various issues in SC and operations management, for example, pricing and inventory management (Ettl et al., 2019), demand prediction (Nguyen et al., 2019), risk management (Zhu et al., 2019), to name a few. More details of data-driven SC applications are reviewed by (Cohen, 2018; Govindan et al., 2018; Misic and Perakis, 2019; Nguyen et al., 2018). Although transportation is one of the key application areas of big data, its use for facility location optimization such as dry ports is still scarce.

Hence, this paper aims to fill the gap by developing a data-driven optimization approach based on data mining and complex network theory to provide practical solutions for the large-scale dry port location problem. The proposed approach has two stages, and is called as the Association Rule Mining with Eigenvector Centrality – Gravity based Community Structure (ARMEC-GCS). In the first stage, we mine a large amount of international transaction data using the ARMEC model to weight the importance of inland regions based on the microeconomic and business perspectives of international customers. In the second stage, the weighting score is then integrated with other factors from macroeconomic (i.e., inland region's foreign trades) and geographic (i.e., spatial distances between inland regions) perspectives in the GCS algorithm to optimize the location and coverage area of dry ports. The ARMEC-GCS approach is validated using the real case of China's Belt Road Initiatives (BRI).

The paper structure is the following. Section 2 reviews the related literature on dry port locations. Section 3 describes the proposed ARMEC-GCS methodology. Section 4 validates the model through the case setting, result analysis, discussion, and robustness checking, as well as highlighting managerial implications. The conclusion and future research directions are in section 5.

2. Review of dry port location studies

The current literature on dry port location analysis is summarized in Table 1. In general, research on dry port locations can be classified into two fundamental design perspectives: microeconomic and macroeconomic. This classification is related not only to the scope of the problem but also to the modeling method adopted.

Table 1: Summary of literature on dry port location problem

	Location selection perspective		Problem size*		Method
	Micro-economic	Macro-economic	Small-scale	Large-scale	
Feng et al. (2013)	✓		✓		Genetic Algorithm
Wang, Chen, et al. (2018)	✓		✓		MIP
Wei and Sheng (2017)	✓		✓		MIP
Ng and Gujar (2009)	✓		✓		Spatial analysis; MIP
Tsao and Thanh (2019)	✓		✓		MIP; Robust optimization
Zhang et al. (2018)	✓		✓		MIP; Game theory
Komchornrit (2017)		✓	✓		MCDM
Li et al. (2011)		✓	✓		Fuzzy Clustering
Canh and Notteboom (2016)	✓	✓	✓		MCDM
Ka (2011)	✓	✓	✓		AHP; MCDM
Chang et al. (2015)	✓	✓	✓		FCM; MIP
Wei et al. (2018)		✓		✓	PCA; Gravity model
Abbasi and Pishvaee (2018)	✓	✓		✓	AHP; MIP
This paper	✓	✓	✓	✓	Data mining; Complex Network

** "Problem size" refers to the size of the studied dry port network, which is classified as small-scale if the study focuses on the city- or regional-level network and as large-scale if the focus is on the nationwide network.*

In the microeconomic perspective, the designer makes the choice of dry port locations based on the economic benefits to be gained from the improved performance of the transportation and supply chain operations. For example, Feng et al. (2013) optimize the location and allocation of the regional seaport and dry port system with the aim of minimizing the sum of the transportation, dry port set-up, and maintenance costs. The dry port location in Wang, Chen, et al. (2018) is selected, taking into

consideration the transportation cost and the cost of opening/closing new/existing facilities. Wei and Sheng (2017) and Ng and Gujar (2009) also choose cost savings in logistics as the primary objective in their dry port location models. Zhang et al. (2018) optimize the dry port locations and pricing strategy for profit optimization. Tsao and Thanh (2019) optimize the sustainable dry port network design which minimize the economic, environment and social costs. All these studies formulate the location optimization problem as a compact MIP model, where the optimal dry ports are selected only from a fixed set of candidate locations given in advance. Another concern is that the optimal solution may only hold true to the specific network topologies used for its model development. As a result, these simplifying assumptions seem to constrain the discovery of the truly optimal location and the practical application of the findings (Zheng et al., 2018).

On the other hand, many researchers take a broader macroeconomic perspective in which dry port locations are considered a multi-criteria decision, allowing conflicting objectives from various stakeholders to be taken into account. For instance, transportation condition, local policy environment, and regional economic development are among the common evaluation indicators for dry port locations (Canh and Notteboom, 2016; Chang et al., 2015; Ka, 2011; Komchornrit, 2017; Li et al., 2011; Wei et al., 2018). Most traditional multi-attribute methods have been adapted to dry port locations, including the analytical hierarchy process (AHP) (Abbasi and Pishvaee, 2018; Ka, 2011), MCDM (Canh and Notteboom, 2016; Komchornrit, 2017), and fuzzy clustering (Chang et al., 2015; Li et al., 2011). However, one of the major drawbacks of these methods is that the weight ranking and decision rules of the location criteria are assessed according to human perception and experience, which are more or less biased, subjective, and difficult to quantify accurately (Canh and Notteboom, 2016). Another common concern is that the locations derived from the multi-attribute decision making are typically optimal at the macro level only, while from a microeconomic and operational perspective, there is no guarantee they would be able to attract sufficient demand from shippers to stay economically viable (Chang et al., 2015; Liu et al., 2018).

Some researchers are also attempting to adopt both the microeconomic and macroeconomic perspectives to complement the way they limit each other, by developing a two-stage dry port location optimization approach. As such, a set of candidate locations is first selected using the multi-criteria model at the macro level. Then from the candidate set, a MIP model is performed to select the final dry port location that can optimize the performance of the logistics network at the microeconomic and operational levels (Abbasi and Pishvaee, 2018; Chang et al., 2015).

Regarding problem size, when using a conventional location modelling approach such as MCDM and MIP, most existing models for dry port location can only address the small-scale optimization problem specific to the city- and regional-level transportation systems. Thus, the large-scale dry port location problem at the nationwide level has been largely overlooked. In fact, we only found two papers in the current literature that discuss national dry port development (Abbasi and Pishvaee, 2018; Wei et al.,

2018). However, the optimal locations they obtained still suffered from being highly subjective and biased, due to the use of MCDM for the location criteria ranking, as explained above.

In summary, our literature review reveals the absence of a method that can effectively and unbiasedly optimize the large-scale dry port location problem, taking into account both macro- and microeconomic design perspectives. Hence, our proposed data-driven approach, the ARMEC-GCS, which combines nonparametric, scalable algorithms from the data mining domain and complex network theory, can address the gap effectively.

3. Methodology

The overview of the proposed two-stage ARMEC-GCS approach is shown in Figure 1, while the detail of each stage is described in the following subsections.

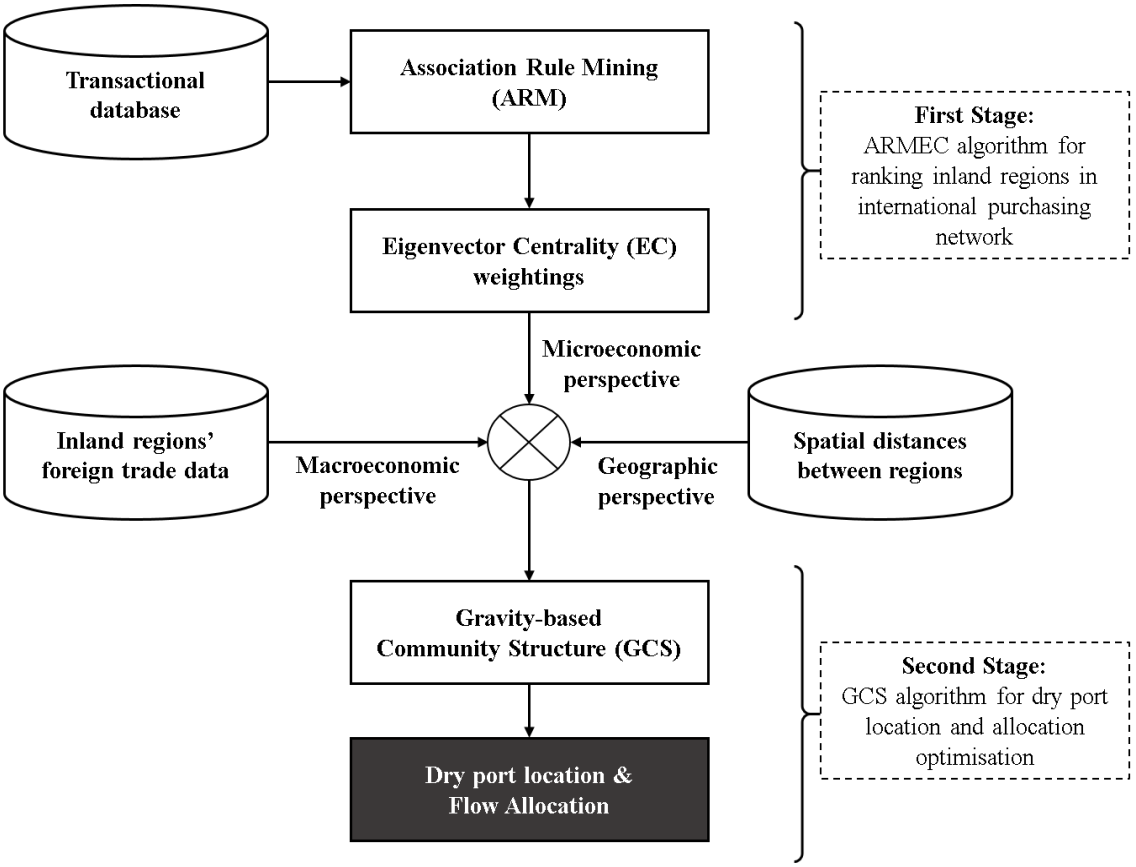


Figure 1 The ARMEC-GCS approach

3.1. Constructing international transaction database

To identify optimal inland cities for locating dry ports based on their trading attractiveness to international customer demand, we construct a large-scale international transaction database recording all demand and supply information, for example, product type, transaction value, buyer location, supplier location, order date, supplier's company size, reputation, production capacity, trading

capability, etc. Nevertheless, the focus of the present study is to discover association rules between international demand patterns and supplying locations; therefore, we only select buyer- and product-related attributes. In particular, the international demand pattern is represented by a matrix of $D(x_i, y_i, z_i)$, in which each matrix element accommodates a key feature of the i^{th} transaction, including the buyer location (x_i), the production lead time (y_i), and the transaction value (z_i). International demand patterns are distinguished by the unique combinations of these three features. The demand matrix, together with the supplier location (s_i), forms the transactional database used for the ARM model. A sample of the international transactional database can be seen in Table 2.

Table 2: Example of the international transactional database

Supplier ID	Transaction ID	Buyer location (x)	Lead time (y)	Transaction value (z)	Supplier location (s)
A1	TID1	Poland	7 days	Low*	Nanjing
A1	TID2	India	30 days	Very high	Nanjing
A2	TID4	Romania	60 days	Medium	Foshan
A2	TID5	Finland	20 days	Low	Foshan
….	…	….	…	….	…

**: is ranked by Alibaba based on the transaction level*

3.2. Stage 1: ARMEC algorithm

3.2.1. Association rule mining (ARM)

Data mining is the process of applying a wide range of machine learning and statistical techniques in order to extract previously unknown patterns for better decision making (Corne et al., 2012). ARM is among the most versatile and widely used data mining techniques (Nguyen et al., 2018). It is the method of finding frequent patterns, associations, co-occurrences, or causalities between a complex set of attributes in big data (Ting et al., 2014). Such rules have been well-adapted to support various decision making, for instance, new product development (Bae and Kim, 2011), logistics quality control (Ting et al., 2014), and fraud detection in procurement management (Ghedini Ralha and Sarmento Silva, 2012). ARM has also been used to optimize location-related problems, such as shelf-space allocation (Tsai and Huang, 2015), storage assignments (Chiang et al., 2011), and logistics scheduling (Lee, 2016), which is relevant to our studied problem of dry port location.

The output of the ARM is a set of association rules that can be expressed in the format {A} => {B}, where A and B refer to the antecedent and consequence part of the rule, respectively. In this study, the ARM aims to evaluate the supplying capability and trading attractiveness of inland regions from the business perspective of international customers. Thus, we only focus on association rules for which the antecedent (A) is the set of international demand patterns and the consequence (B) is the set of suggested supplying locations.

There are many measures of rule strength or importance, as explained in De La Iglesia et al. (2006). In this paper, we use the most common ones, namely *support* and *confidence*. The rule support refers to the probability that both the antecedent and consequent occur together, while the rule confidence is the conditional probability that the consequence occurred based on the occurrence of the antecedent (Padmanabhan and Tuzhilin, 2003). While the support implies the coverage (or frequency) of the rule in the transaction database, the confidence indicates the rule strength (or reliability) (Witten and Frank, 2011). Typically, a rule is considered as important and interesting if it satisfies both the minimum support and minimum confidence thresholds predefined by domain experts. The mathematical expression of support and confidence is as follows:

$$Support = P(A \cap B) = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions} \quad (1)$$

$$Confidence = \frac{P(A \cap B)}{P(A)} = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions\ with\ A} \quad (2)$$

Given the fact that the number of rules grows exponentially, which makes the brute-force approach infeasible, this paper thus adopts one of the most popular ARM algorithms, called Apriori (Ghedini Ralha and Sarmento Silva, 2012). The Apriori algorithm involves two stages. In the first stage, it performs a breadth-first search to generate a large set of candidate itemsets from which frequent itemsets are identified. The principle here is that an itemset is considered a frequent itemset if all of its subsets have support higher than the predefined minimum support threshold. In the second stage, the identified frequent itemsets are then used to generate association rules. Similarly, only rules that have confidence higher than the predefined minimum confidence threshold are considered interesting and worth further analysis.

3.2.2. Eigenvector centrality (EC) in complex network theory

Parallel to rapid progress in studying big data analytics, another emerging research stream is big data visualization, which involves multiple techniques to make the result of big data analytics more understandable, accessible, and useable for timely data-driven decision making (Nguyen et al., 2018). Among different visualization techniques, complex network analysis has been proven one of the most scalable techniques for dealing with large, complex data. Unlike classical network theory, the complex network focuses primarily on studying the nontrivial topological patterns that are neither uniformly ordered nor random (Rubinov and Sporns, 2010). Since such complex patterns are inherently linked to most real-world systems, the method has gained much attention from a wide range of research fields, such as biology (Rubinov and Sporns, 2010), transportation (Saberi et al., 2017), and social networks (Verma et al., 2018). There are a number of measures to describe the structural properties of a complex network. In this paper, we use two fundamental measures: EC for the ARMEC model and community structure for the GCS model (see section 3.3).

After generating a set of association rules by using the Apriori algorithm described in section 3.2.1, the next step in the ARMEC model is to develop an international purchasing network in which all objects and relationships among the association rules are represented as nodes and edges. In such a network, nodes include association rules, their associated antecedents (i.e., international demand patterns) and their associated consequences (i.e., supplying locations). Causal relationships among the association rules are illustrated by directed edges. An example of this network can be seen in Figure 2a. However, within the scope of this study, we focus particularly on the nodes representing the supplying locations; therefore, the network excludes nodes representing demand patterns, as seen in Figure 2b. In network (b), the size of the red node represents the strength of the association rule measured by its confidence value, whereas the size of the green node indicates the centrality of the supplying location in the network.

In network analysis, node centrality refers to the importance of a node in the network. There are various indices to measure node centrality, including degree, closeness, eigenvector, clustering coefficient, betweenness, and information index (Wang, Li, et al., 2018). In this paper, we use eigenvector to measure the centrality in the international purchasing network (Figure 2b).

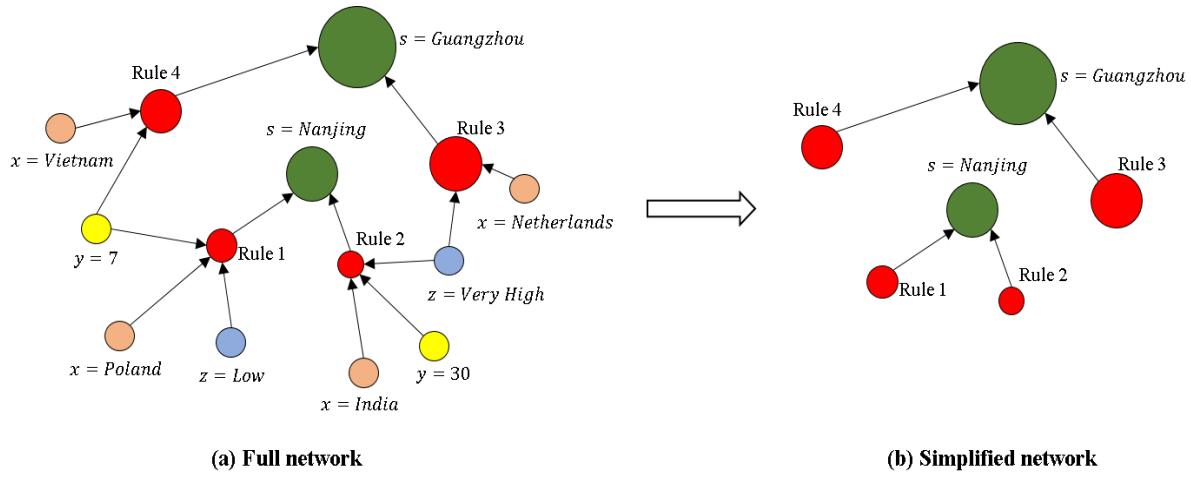


Figure 2 Example of international purchasing network used in this study

Red node - Association rule ID; Green node - Supplying location; Orange node - Buyer location; Yellow nodes - Product type based on production lead time (days); Blue nodes - Transaction value

The definition of eigenvector centrality (EC) in this study is adopted from Ghanbari et al. (2018). Let $G = (V, E)$ be the network G containing a set of nodes V and a set of edges E. The network can be represented through its adjacency matrix $A = \{A_{ij}\}$, where A_{ij} is a binary variable that takes 1 if an edge exists between node v_i and node v_j otherwise 0. EC of a node is determined by the number of its connected neighbors and the importance of each neighbor. Hence, the EC k_i of node v_i is proportional to the sum of the centralities of its connected neighbors. As such:

$$k_i = \frac{1}{\lambda}\sum_{1}^{n} A_{ij}x_j \quad (3)$$

where n is number of neighbors linked with node v_i, x_j is the value of the neighbor node v_j, and λ is the largest eigenvector value in the adjacency matrix A.

Compared to other centrality indices such as degree centrality which measures the node importance simply by counting number of neighbours connected with the node, many studies of transportation networks favor EC since it can provide more profound insights about the node influence in the network (Brookes and Huynh, 2018; El-adaway et al., 2018; Parajuli and Haynes, 2018).

3.3. Stage 2: GCS algorithm

The dry port-based inland transportation system could be regarded as a complex network where the ports could be represented by nodes and the relationship among ports could be represented by edges. To determine dry port locations and their coverage areas in such a network, we adopt the concept of community structure from complex network theory. Community structure (so-called clusters or modules) is a common phenomenon in many real-world networks, referring to partition of a network into groups (or communities) of nodes which are densely connected within the groups and sparser connected with nodes in other groups (Costa, 2015). Several studies have been published using community structure theory in the transportation and logistics research area such as cargo ship movement analysis (Kaluza et al., 2010), global logistic network design (Sun et al., 2012) and global hub location optimization (Zheng et al., 2018). In general, it is feasible to use community structure theory to detect port relationships at a large-scale network level.

A range of approaches have been developed to detect the community structure in complex networks, for example, spectral-based, clustered-based, and modularity-based algorithms (Zhou et al., 2018). Among these, the modularity-based algorithm has been widely applied in large-scale networks, due to its fast, efficient computation (Clauset et al., 2004). Modularity is a quality function to measure whether a particular partition of the network into communities is good, in the sense that there is a high density of edges within communities and only sparse connections between them. Newman and Girvan (2004) define modularity (Q) as follows.

$$Q = \sum_{i}(e_{ii} - {a_i}^2) \quad (4)$$

where e_{ii} equals to the fraction of edges that connect vertices within community i. It is the main diagonal elements of the symmetric matrix $E = \{e_{ij}\}$, where element e_{ij} is the fraction of edges in the network that connect vertices in community i to vertices in community j. The mathematical expression of e_{ij} is given by Clauset et al. (2004) as follows:

$$e_{ij} = \frac{1}{2m}\sum_{uv} A_{uv}\ \delta(c_u, i)\delta(c_v, j) \tag{5}$$

where A_{uv} is an element of the adjacency matrix, which takes 1 if vertex u and vertex v are connected, and 0 otherwise; m is the total number of edges in the network, measured by $\frac{1}{2}\sum_{uv} A_{uv}$. If vertex u belongs to community i, then $\delta(c_u, i)$ equals to 1, and -1 otherwise. Similarly, if vertex v belongs to community j, then $\delta(c_v, j)$ equals to 1, and -1 otherwise.

Furthermore, ${a_i}^2$ in Eq (4) is the expected fraction of edges that connect to vertices in community i when the end of edges are connected at random. The expression of a_i is formulated in Clauset et al. (2004) as follows:

$$a_i = \frac{1}{2m}\sum_{i} d_u \delta(c_u, i) \tag{6}$$

Where d_u is the degree centrality of vertex u, measured by $d_u = \sum_1^n A_{uv}$.

Here, the modularity-based community detection model becomes a mixed-integer quadratic programming problem of which the objective is to find the optimal splitting point of the network to maximize the modularity in Eq. (4). Previous studies have addressed the modularity maximization using both exact (eg. Costa 2015) and heuristic approach (eg. Santiago and Lamb 2017). However, when dealing with large-scale, real-world facility location problems, using approximate optimization techniques such as greedy heuristic is an ideal choice to effectively search over a large feasibility space for optimal solutions (Ishfaq and Sox, 2011; Ruiz et al., 2018; Santiago and Lamb, 2017). Therefore, to optimize the location and service area of dry ports, this paper employs one of the most widely used algorithms in the modularity-based community structure theory, called the fast Newman (FN) algorithm (Newman and Girvan, 2004). It adopts an agglomerative approach to search the optimal network splitting points in a greedy manner.

However, the classical FN algorithm was developed specifically for an unweighted network, while the dry port transportation system is typically a weighted network of which edge weights indicate the logistics relationships between nodes. Hence, in this paper, we adopt the improved FN algorithm which can also be used for the weighted network (Liu et al., 2013; Newman, 2004a; Zhang and Meng, 2019). In particular, e_{ij} in Eq. (5) and a_i in Eq. (6) are redefined as:

$$e_{ij} = \frac{1}{2w}\sum_{uv} W_{uv}\ \delta(c_u, i)\delta(c_v, j) \tag{7}$$

$$a_i = \frac{1}{2w}\sum_{u} W_u \delta(c_u, i) \tag{8}$$

where W_{uv} is the edge weight between vertex u and vertex v; W_u is the vertex weighted degree, which equals to the summation of edge weight attaching to vertex u; and w is the summation of edge weight in the network, measured by $\frac{1}{2}\sum_{uv} W_{uv}$.

In this study, the edge weight, which indicates the logistics relationship between two locations, is measured using the gravity model. Based on Newton's universal law of gravity, the gravity model provides a realistic, applicable tool to describe and predict the interaction between objects, taking into account both their mass and spatial characteristics (Campbell and O'Kelly, 2012). The model has been widely applied to international trading networks, logistics hub locations, and in many other social science research fields (Anderson and van Wincoop, 2003; khosravi and Akbari Jokar, 2017; Zeng et al., 2017; Zhang and Meng, 2019). In this study, the gravity model is extended to measure the logistics relationships among inland regions, based on their spatial characteristics and logistic quality from both the macroeconomic and microeconomic perspectives. The extended gravity function measuring the edge weight W_{uv} between region (vertex) i and j in the dry port network is expressed as follows:

$$W_{uv} = \frac{T_u\, T_v}{{D_{uv}}^2\,(1 - Z_{uv})^2} \tag{9}$$

where D_{uv} is the spatial distance between regions u and region v. T_u, T_v are the logistics quality of regions u and region v from the macroeconomic perspective. Since the main function of dry ports is to improve the connectivity between inland regions and international gateways (eg. Seaports or cross-border train stations) for increased international trading, T_u, T_v can be measured by the total value of import and export trade through regions i and j, respectively. Prior literature has adopted such foreign trade values as evaluative criteria for dry port locations at the macro level (Chang et al., 2015; Li et al., 2011; Wei et al., 2018). Finally, Z_{uv} is the gravity coefficient adopted in the gravity function to represent the external force affecting the logistics interaction between two regions. As discussed above, the ARMEC model distinguishes the difference between regions by their EC scores, which weight the importance of regions in the international purchasing network. Since the EC score of a region depends critically on its associations with purchasing patterns of international customers, it can be used to represent the logistic quality of regions from microeconomic and business perspectives. Thus, the gravity coefficient Z_{uv} can be calculated by:

$$Z_{uv} = k_u k_v \tag{10}$$

where k_u, k_v are the EC scores of regions u and region v, respectively, obtained by Eq. (3) in the ARMEC at stage 1.

From all the adjustments above, the classical FN algorithm is elaborated to fit the weighted network of dry ports in our study. We call the new algorithm the gravity-based community structure (GCS). The main steps of the GCS algorithm are as follows:

Step 1: Network initialization: Convert the studied geographical area into an unweighted network with nodes (cities) and edges.

Step 2: Converting the unweighted network into the weighted network by calculating the edge weight W_{uv} between any pair of nodes, using Eq. (9). In this network, each node is treated as one community.

Step 3: Community combination.

- Sequentially join any two communities together and calculate the modularity variation ΔQ. Based on (Newman, 2004b), ΔQ is computed by:

$$\Delta Q = e_{ij} + e_{ji} - 2\,a_i a_j = 2(e_{ij} - a_i a_j) \tag{11}$$

  where e_{ij}, a_i, and a_j are obtained using Eq. (7) and Eq. (8).

- On the basis of the greedy algorithm, select the join that results in the maximum increase or minimum decrease in modularity. The modularity of the new communities is computed.

Step 4: Update the elements e_{ij}.

Step 5: Execute Step 3 and 4 repetitively until the whole network is merged into one community.

Step 6: The best division is selected with the highest modularity in the process. As a result, the network is split into a set of communities. In each community, the vertex with highest weight (most influential) is selected to locate a dry port hub, fed by other vertices within the same community. The weight (r_u) of vertex u is calculated as follows:

$$r_u = \sum\nolimits_v W_{uv} \tag{12}$$

where W_{uv} is the weight of the edge having connection to vertex u, measured by Eq. (9).

4. Experiment and model validation

In this section, we apply the proposed ARMEC-GCS approach to find optimal locations of dry ports and allocations of their service areas in Mainland China in the context of the BRI framework. China is chosen as the case application in this study given the fact that the country has recently initiated a large number of dry port development projects as the key enabler to reach its full international trade growth potential (Wei et al., 2018; Xie et al., 2017).

4.1. Case study: Dry port developments under China's Belt and Road initiative (BRI)

In 2013, China launched the BRI to enhance the infrastructure connectivity between Asia, Europe and Africa, laying a stronger foundation for international trade and regional economic growth (Huang, 2016). Since then, the BRI has become one of the world's largest infrastructure and investment projects in history, with the participation of 65 countries, accounting for 63% of the world population and 30%

of the global gross domestic product (Sarker et al., 2018). It is estimated that the total investment in BRI projects will reach up to USD 7.4 trillion by 2030, and more than 80% of which will be used for infrastructure developments of two mega projects: the Belt and the Road (Swiss Re Institute, 2017). The "Belt" refers to the "Silk Road Economic Belt" (SREB), comprising six international overland economic corridors connecting China with Central Asia, West Asia, the Middle East, and Europe. The "Road" refers to the sea routes called the "21st Century Maritime Silk Road" (MSR), linking the South China Sea, the South Pacific Ocean, and the Indian Ocean (Chen et al., 2018). The geographical coverage of the BRI is depicted in Figure 3.

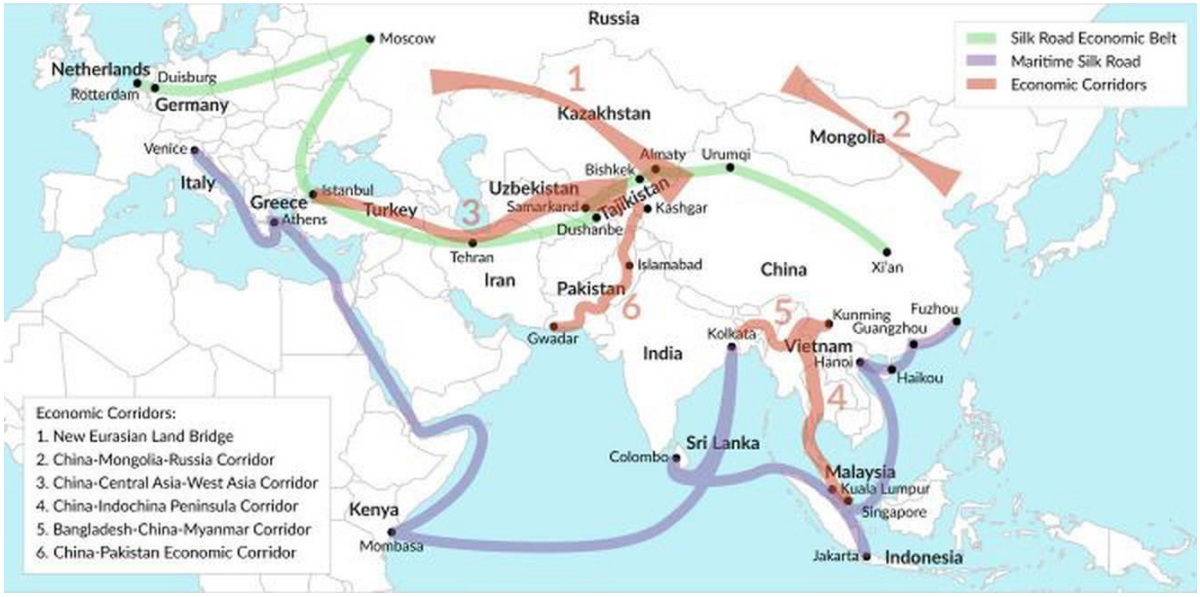


Figure 3 The Belt and Road framework

A recent report by Konings (2018) claims that in the long run, the improvement in transport facility will halve overall trade costs between the BRI countries and will increase their cross-border trade by 35%–45%. Under such circumstances, using dry ports to ease congestion at port gateways and improve inland access is particularly essential to guarantee the efficiency of the entire transportation chain (Yu, Fransoo, et al., 2018). In fact, dry ports have been set to play an integral part in the future implementation of the BRI framework, as stated by the Ministry of Transport of the People's Republic of China (2017). However, the Ministry also described the current development of dry ports in Mainland China as "blind constructions" with a lack of unified strategic planning. Hence, this experiment aims to test whether our proposed ARMEC-GCS approach can provide a valid and applicable solution for the large-scale problem of dry port locations in China.

In particular, we aim to find optimal dry port locations and their allocated service areas to cover all 309 prefecture cities in Mainland China, apart from those like Qinghai, Tibet and Guizhou Province without a dry port operation in place (Wei et al., 2018). These studied inland cities come from 24 inland

provinces, namely Sichuan, Anhui, Fujian, Gansu, Guangdong, Guangxi, Hainan, Hebei, Heilongjiang, Henan, Hubei, Hunan, Inner Mongolia, Jiangsu, Jiangxi, Jilin, Liaoning, Ningxia, Shaanxi, Shandong, Shanxi, Xinjiang, Yunnan, Zhejiang. The location problem investigating up to 95,481 edges among 309 city nodes is one of the largest-scale networks in the dry port location literature, which demonstrates the real need to use scalable solution approach such as the ARMEC-GCS.

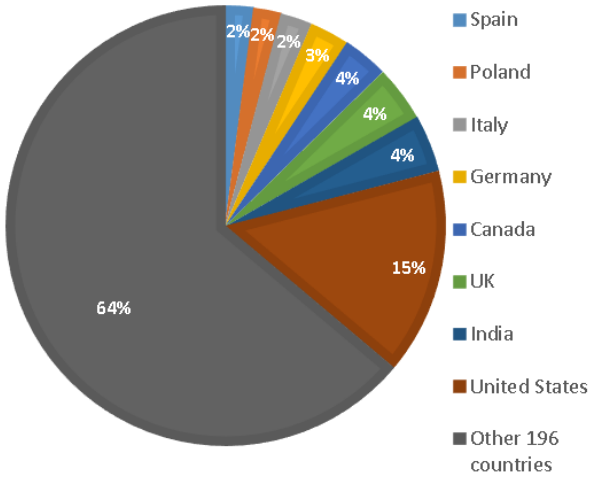
4.2. Data collection

4.2.1. Data collection for stage 1

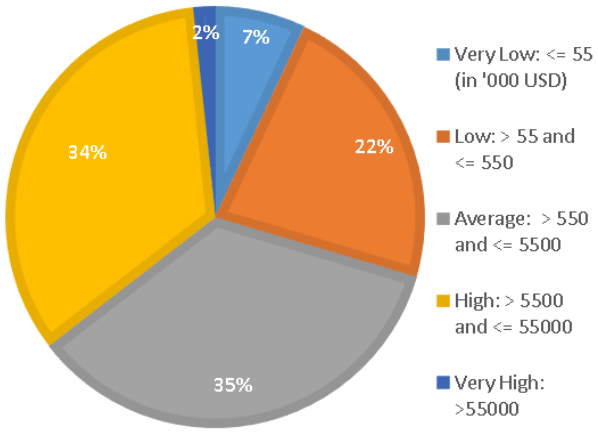
In the first stage of the ARMEC-GCS approach which extracts insights between international demand patterns and Chinese suppliers, we construct a large transactional database from Alibaba.com. Alibaba is chosen not only because it is the world's biggest data source for business-to-business international trading, covering over 200 countries and regions, but also due to its pivotal role in the development of the Digital Silk Road as part of the BRI framework (Silin et al., 2017). In fact, Alibaba is currently developing 14 data centers around the globe, equipped with a 5G communication network, with the aim of supporting goods movement and unifying custom procedures among 10 countries along the SREB (Silin et al., 2017).

We use a web crawler to collect supplier information and sales transaction records from Alibaba.com. Since one of China's main economic interests in the BRI is to boost its inland regions towards an export-oriented economy (Huang, 2016; Wei et al., 2018), we only collect data from Chinese suppliers who provide international shipping routes across countries within the BRI projects. The transaction data we collect in this study include machinery and equipment, as they account for more than 50% of total China exports to the EU (Konings, 2018).

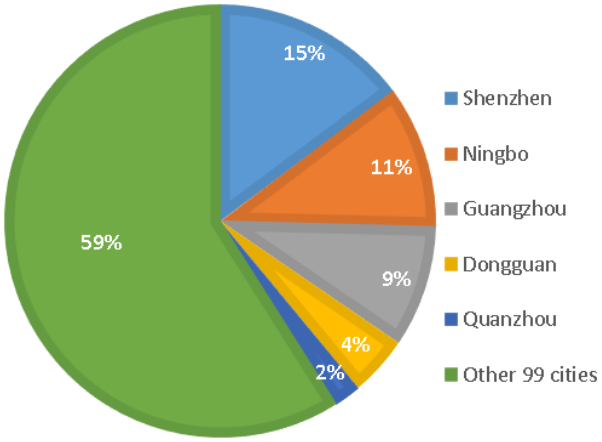
As a result, our crawler returns two separate datasets. The first dataset contains supplier information, while the second provides the whole transaction history of each supplier. These datasets can be joined together for data mining through the suppliers' unique IDs. After removing missing data, excluding domestic transactions and joining the two datasets, our joint dataset includes 25,643 transactions between China and international customers. Each transaction is featured by 45 attributes from the buyer and supplier. Numerous data are stored, but not all can be used to model international demand patterns. As described in section 3.1, we represent an international demand pattern through a matrix of $D(x_i, y_i, z_i)$, a compound of the transaction's buyer location (x_i), production lead time (y_i) and transaction value (z_i). The demand matrix is then mined by the ARM to find associations with the Chinese supplying location (s_i). The overall description of the Alibaba international transaction database used in this experiment can be seen in Figure 4.



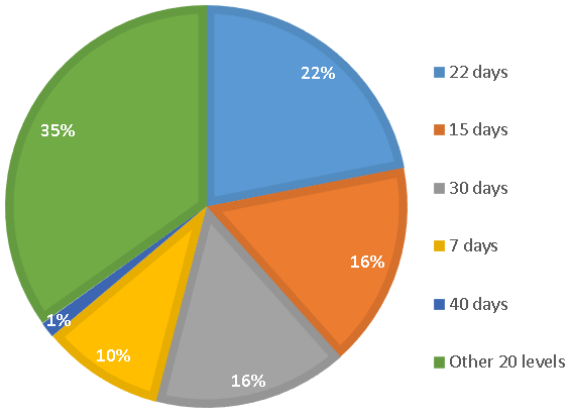
Buyer location
(204 countries)



Transaction value
(5 levels)



Supplier location
(104 Chinese cities)



Production lead time
(25 levels)

Figure 4 Overall description of the Alibaba international transaction database.
(25,643 total transactions)

4.2.2.Data collection for stage 2

In the second stage, we apply the GCS algorithm, as described in section 3.3, to find the optimal locations for dry ports as well as to determine the coverage area of each dry port. As explained in Eq. 9, the input data for the extended gravity model to measure the logistic relationships (i.e., edge weight, W_{uv}) between inland cities includes: (1) The gravity coefficient (Z_{uv}) based on the EC scores (k_u, k_v) of each city obtained from the ARMEC stage; (2) The logistics quality (T_u, T_v) of each city measured by the total import and export value obtained from its 2016 Statistical Yearbook; and (3) The spatial distance (D_{uv}) between each pair of city nodes, measured in miles, based on their longitude and latitude coordinates.

4.3. Experiment results and discussion

4.3.1. Stage 1 - ARMEC

- **ARM results**

Based on the constructed Alibaba database described in Section 4.2.1, we perform the Apriori algorithm in the R program to extract the association rules between international demand patterns (antecedent) and Chinese supplying locations (consequence). Regarding the minimum support and minimum confidence thresholds, many studies tend to set them at relatively high values to limit the number of rules generated, and decision making is derived only based on the top rules with the highest support and confidence (Ting et al., 2014). However, in order to evaluate the scalability of our proposed approach, this experiment is conducted with very low minimum support and confidence thresholds, to ensure no important rules are missed out. Since the lowest occurrence frequency for itemsets in our Alibaba dataset is 0.000037, it is reasonable to set the minimum support threshold equal to 0.000037. As the transaction data in this paper are sparse, the value of the minimum confidence threshold is set at its first quantile of 0.4 (40%) to avoid over-pruning informative rules while ensuring the trivial rules are excluded, as suggested by Belyi et al (2016). As a result, a total of 3,110 association rules are generated, and these international demand patterns (antecedent) are satisfied by 80 inland Chinese cities (consequence). Table 3 provides the statistical summary for international demand patterns within these rules. Examples of the top 10 rules with the highest confidence can be seen in Table 4.

Table 3: Statistical description of the distribution for 3,110 rules

Antecedent size	Number of rules	Support			Confidence		
		Min	Mean	Max	Min	Mean	Max
1-itemset	81	0.00004	0.00040	0.00370	0.4	0.69775	1
2-itemset	1419	0.00004	0.00018	0.01778	0.4	0.69485	1
3-itemset	1610	0.00004	0.00004	0.00312	0.4	0.67212	1

Table 4: Top 10 out of 3,110 rules sorted by confidence

Antecedent of the rule	Consequence of the rule	Support	Confidence
{ x =Niger}	=> {s = Zhongshan}	0.000039	1
{ x =Jersey}	=> {s = Zhangzhou}	0.000039	1
{ x =Indonesia, y =50}	=> {s = Tangshan}	0.000273	1
{ x =United Kingdom, y =4}	=> {s = Chengdu}	0.000117	1
{ x =Luxembourg, z = Very high }	=> {s = Quanzhou}	0.000195	1
{ x =Switzerland, y =3}	=> {s = Ningbo}	0.000078	1
{ x =Afghanistan, z = Very high }	=> {s = Foshan}	0.000156	1
{ x =Austria, y =15, z = High }	=> {s = Anqing}	0.000078	1
{ x =South Africa, y =30, z = Low}	=> {s = Jinzhou}	0.000039	1
{ x =Iceland, y =20, z = Average}	=> {s = Shantou}	0.000156	1

- **EC-based importance of Chinese cities in international purchasing network**

While the previous section determines a set of frequent rules in general, this section will demonstrate the advantage of our approach, which uses a complex network to deal with the large-scale, complex relationships among these rules. In particular, all 3,110 association rules found can be visualized as a network, using popular software called Gephi. Since our main focus is on the Chinese supplier locations, Figure 5 displays the network that describes the relationship among the 3,110 rules (red nodes) and their associated consequences of 80 Chinese supplying cities (green nodes). The size of the red nodes represents the strength of the association rules, measured by their confidence values, whereas the size of the green nodes indicates the influence of Chinese inland cities, measured by their EC score analysis. The full list of EC scores for 80 cities is provided in Table A in Supplemental Material.

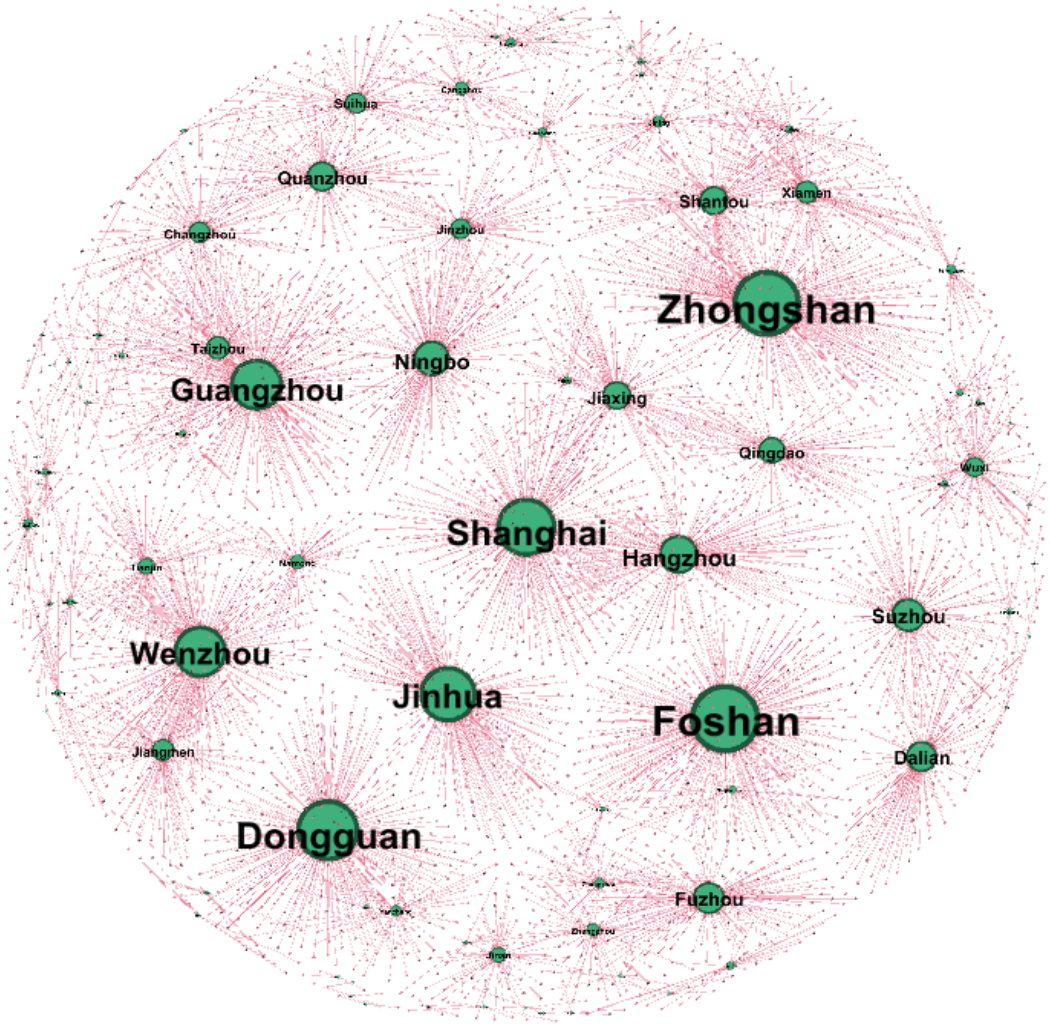


Figure 5 An international purchasing network that visualises 3,110 association rules (red nodes) and their associated Chinese supplying cities (green nodes).

4.3.2.Stage 2 – GSC algorithm

With the input data described in section 4.2.2, the GCS algorithm in the second stage is run on the MATLAB program. According to the final result returned from the GCS algorithm, 309 inland Chinese cities from 24 provinces are grouped into 13 communities, in which each community is served by a hub dry port with the highest degree of centrality (i.e., the most influential) in the community. The suggested dry port locations and their coverage areas are presented in Figure 6. For the managerial discussion, we also include in the figure the locations of major seaports and international train gateways under the BRI framework. The map codes can be seen in Tables 7 and 8.

Figure 6 Dry port locations and their coverage areas by the ARMEC-GCS approach
(Node codes are shown in Tables 5 and 6)

The optimal dry port locations pinpointed by the ARMC-GCS approach are also closely in line with the real BRI development plan. Among 13 optimal locations, some already has the ongoing BRI dry port development projects such as Shenyang, Xi'an, Chaozhou and Xingtai, while the others currently serve as the BRI international gateways such as Beijing, Urumqi, Chengdu, Guangzhou, Suzhou, Yantai, and Xiangtan.

Moreover, the ARMC-GCS approach is also credible in terms of capturing real spatial characteristics when detecting the distinctive community structure of Community 13 (Bayingolin Mongol

Autonomous Prefecture) and Community 2 (Jiayuguan and Jiuquan). Indeed, these two communities have demographic mechanisms different from other subdivisions in Xinjiang and Gansu provinces.

The role of each suggested dry port location in the BRI's actual development plan is highlighted in Table 7. Since the optimal solutions include the key transportation hubs which closely reflect the real BRI development plan, the ARMEC-GCS approach is validated.

4.4. Robustness check

In this section, we include two tests to check the robustness of the proposed ARMEC-GCS method's performance.

4.4.1. Test 1: Comparing the ARMEC-GCS approach and the GCS-only approach

In this test, the solution for dry port locations and their assigned coverage areas is derived based on the GCS-only approach, meaning without using the ARMEC model to mine association rules between international demand patterns and supplying locations. The results can be seen in Figure 7 and Table 5.

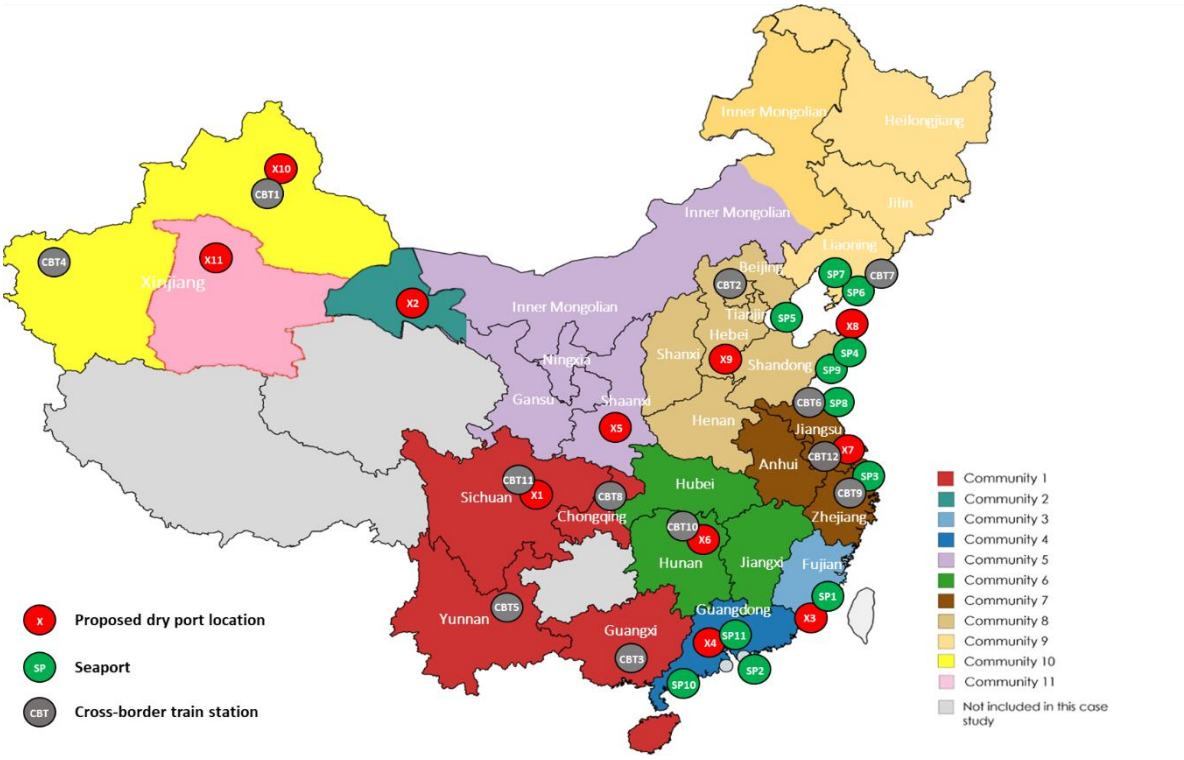


Figure 7 Dry port locations based on the results of the GCS-only approach.
(Node codes are demonstrated in Tables 6 and 7)

Table 5: Dry port locations with assigned communities using the GCS-only approach

Dry port code	Hub dry port location	Community size (Number of nodes)
X1	Chengdu	51
X2	Jiayuguan	2
X3	Xiamen	14
X4	Guangzhou	19
X5	Xi'an	39
X6	Xiangtan	32
X7	Suzhou	33
X8	Yantai	48
X9	Xingtai	57
X10	Urumqi	13
X11	Korla	1

As compared to the ARMEC-GCS, the GCS-only algorithm fails to suggest some key dry port locations, such as Beijing and Shenyang. As mentioned above, these two nodes are key nodes of the CMREC and the MSR. Without these dry port nodes, the wide inland areas of the Inner Mongolian, Northern, and Metropolitan areas of China would easily fall into the disorder of logistics operations as the increasing volume of hinterland cargo from and to seaports will lead directly to severe traffic congestion, longer shipping times, and shortages of capacity at the seaports. The whole global shipping service would also suffer.

4.4.2.Test 2: Sensitivity analysis

In our ARMEC-GCS approach, two main parameters have impacts on the result: the minimum support and minimum confidence thresholds in the ARM model. Thus, we conducted the sensitivity analysis by examining how the results would change when the values of these two parameters changed. The results can be seen in Table 6.

Table 6: Sensitivity analysis result

Case	Support	Confidence	Number of rules	Number of cities
Case 1	0.000030	0.4	3110	80
Case 2	0.000030	0.7	1908	71
Case 3	0.000030	1	1837	71
Case 4	0.000100	0.4	737	53
Case 5	0.000100	0.7	353	44
Case 6	0.000100	1	186	37
Case 7	0.001000	0.4	51	16
Case 8	0.001000	0.7	13	10
Case 9	0.001000	1	5	5

As shown in Table 6, the number of association rules and number of cities ranked in the ARMEC model are quite sensitive to different settings of minimum support and minimum confidence thresholds. As the GCS model takes the ARMEC output as its input, the optimal locations and service areas of dry ports are also likely to change accordingly.

In particular, when increasing the support and confidence thresholds, the number of rules drops significantly, which means less computation resources required. However, the number of cities also considerably reduces, which indicates information lost. Having a closer look at the dry port locations, we notice that Beijing and Shenyang only appear in the ARMEC results in case 1 (Table 6) of which the support threshold sets at the lowest value. This is because the Alibaba dataset is quite large and sparse, which is common in real-world data. Therefore, the minimum support threshold needs to be as low as possible to avoid losing important information. However, by doing so, it will result in a large number of association rules (in this case, 3110 rules as depicted in Figure 5), which require high computational cost to deal with. Thus, it is essential to develop a scalable visualization model that is capable of analysing the big set of association rules effectively. For that, the use of an EC-based complex network, as proposed in our ARMEC-GCS approach, is an effective way to enhance the scalability of the whole method.

4.5. Managerial implications

In reality, the current dry port development in China is characterized by blind construction and a lack of strategic planning (The Ministry of Transport of the People's Republic of China, 2017). As a result, more than 100 dry ports have been built to serve the demand of over 300 inland cities. Such excessive construction could lead to overcapacity, a low utilization rate and limited returns on investment (Chang et al., 2015).

Using the ARMC-GCS approach, our results show that the Chinese inland transportation system has a strong community structure since 309 cities can be clustered cohesively into 13 communities. The average community size is quite large, which implies that the suggested dry port location in each community has strong hub functions and can attract sufficient traffic volume to be financially viable.

To help remove the current predicament of the Chinese dry port development, Table 7 provides some actionable insights for port authorities to decide whether or not to set up a new dry port or close the existing one based on its role in the BRI development plan, while Table 8 recommends some potential partnerships between dry ports and seaports/cross-border railways.

Table 7: Dry port locations with assigned communities using the ARMEC-GCS

Optimal solutions in this study	Rationale -

Status	Dry port location name	Number of cities allocated to the dry port	Roles of the locations in the BRI's actual development plan
Existing[5]	Chengdu (DP1)	56	Chengdu is the largest trade center in Western China and also the Asia′s largest rail freight transport hub (Post and Parcel, 2016). One of the three key NELBEC[1] projects is Chengdu – Lodz (Poland) (Yang et al., 2017).
	Jiayuguan (DP2)	2	Jiayuguan is the key transportation hub in Western China for the SREB plan. Especially, it sits one of the three key NELBEC projects, Chongqing – Duisberg (Germany), the one with numerous road and railway connections to transport goods from China to Central Asia and EU (Samaa Digital, 2017).
	Beijing (DP5)	20	Beijing plays the pivotal node in both the MSR and SREB. It has direct access to Port of Tianjin which is the largest seaport in Northern China, serving 11 northern provinces and also Mongolia. It is also the starting point for one of the two major routes in the CMREC[2], namely Beijing - Tianjin - Hebei - Hohnot - Mongolia – Russia (Lehman Brown International Accountants, 2017).
	Xi'an (DP6)	10	Xi'an is a critical node in the BRI because it is the starting point of the New Silk Road. It also serves as transportation, trading and logistics hub connecting Northwest, Eastern, Central, and Southwest regions of China (KPMG China, 2018). Currently, there is a project to build an international dry port in Xi'an (The Ministry of Transport of the People's Republic of China, 2017).
	Shenyang (DP9)	42	Currently, Shenyang already has a dry port that consolidates cargoes from Anshan, Benxi and Tieling; and then, transporting by shuttle trains to Port of Dalian (Chang et al., 2015). Furthermore, Shenyang also lies on one of the two major routes in the CMREC, namely the Dailan - Shenyang - Changchun – Harbin (Lehman Brown International Accountants, 2017).
	Xingtai (DP10)	50	Xingtai serves as a transport hub that connects the Central China with the Eastern and Northern China. Currently, it also has a dry port partnered with Tianjin seaport (The Ministry of Transport of the People's Republic of China, 2017).
	Urumqi (DP12)	13	Urumqi is a key gateway in the SREB with three out of six economic corridors passing through, namely, NELBEC, CCAWAEC[3], and CPEC[4] (Swiss Re Institute, 2017).
Proposed[6]	Guangzhou (DP3)	33	Guangdong province is the key manufacturing hub having the largest export value among all Chinese provinces and municipalities (HKTDC research, 2019), while its capital city, Guangzhou, gains global recognition as the largest seaport in China and among the leading ports in the MSR (China Daily, 2018). Thus, setting up a dry port in Guangzhou to support the increasing freight traffics in the area is beneficial.
	Chaozhou (DP4)	14	In the implementation scheme of Guangdong's participation in the construction of the BRI, Chaozhou port

			is set to play supporting roles to the major seaports in the MSR like Guangzhou and Shenzen (China Daily, 2015).
	Xiangtan (DP7)	31	Xiangtan is an important node in the NELBEC. Indeed, the first China-EU train route in use was the railway starting from Xiangtan to Hamburg (Germany). Operating since 2008, the route has become the showcase for the economic advantages of the SREB-related projects (Railwaypro.com, 2017).
	Suzhou (DP8)	34	About 10% of all of China's exports come from Suzhou, and one of the main China-EU Silk Road route is the rail service from Suzhou to Warsaw (DHL, 2016). Suzhou also has direct connections to three major BRI international gateways in Ningbo, Jinhua and Lianyungang. Ningbo is the busiest seaport in China, and is also an intersection for both SREB and MSR (en.people.cn, 2018). Jinhua is the home of the Yiwu – Madrid international railway line - the longest railway in the world (13,000 km). Lianyungang is among the Chinese busiest seaports and the starting point of the NELBEC to Rotterdam (Sarwar, 2018).
	Yantai (DP11)	3	Yantai is the transport hub in Eastern China's Shandong province. In 2017, it was awarded as one of the most dynamic cities in the BRI (China Daily, 2017). In 2019, it launches a new freight railway to Duisburd, Germany (Belt & Road News, 2019).
	Korla (DP13)	1	Our model detects Community 13 due to its unique geographical position. It covers the Bayingolin autonomous prefecture for Mongol people in the southeast of Xinjiang. This is also the largest prefecture-level division in China. Setting up a dry port in its capital city, Korla, can help connect the local economy in Bayinggolin with the SREB international gateways in Urumqi and Kashgar, thereby boosting its economic growth.

[1] *New Eurasian Land Bridge Economic Corridor (NELBEC)*
[2] *China-Mongolia-Russia Economic Corridor (CMREC)*
[3] *China-Central Asia-West Asia Economic Corridor (CCAWAEC)*
[4] *China-Pakistan Economic Corridor (CPEC)*
[5] *Existing dry ports in the BRI's actual development plan (The Ministry of Transport of the People's Republic of China, 2017)*
[6] *Proposing to develop new dry ports*

Table 8: Major international gateway ports under the BRI and suggested partnerships with dry ports based on the results of this experiment

Function	Code	Actual international gateway	Suggested partnerships with hub dry ports obtained from this study
Seaport (SP)	SP1	Xiamen	DP4, DP7
	SP2	Shenzen	DP3
	SP3	Ningbo	DP8
	SP4	Qingdao	DP10, DP11
	SP5	Tianjin	DP5
	SP6	Dailian	DP9

	SP7	Yingkou	DP9
	SP8	Lianyungang	DP8
	SP9	Rizhao	DP10, DP11
	SP10	Zhanjiang	DP3
	SP11	Guangzhou	DP3
Cross-border train station (CBT)	CBT1	Urumqi	DP12, DP2, DP13
	CBT2	Beijing	DP5, DP10
	CBT3	Nanning	DP3, DP7
	CBT4	Kashgar	DP13, DP12
	CBT5	Kunming	DP1
	CBT6	Lianyungang	DP6, DP8
	CBT7	Shenyang	DP9
	CBT8	Chongqing	DP1, DP6
	CBT9	Jinhua	DP8
	CBT10	Xiangtan	DP7, DP6
	CBT11	Chengdu	DP1, DP6
	CBT12	Suzhou	DP8

5. Conclusion

On the basis of data mining and complex network analysis, this paper proposes a two-stage ARMEC-GCS approach to optimize the location and service area of dry ports in a large-scale inland transportation system. In the first stage, we use ARM to extract, from a large transaction database, a set of association rules between international demand patterns and supplying locations. These association rules are then visualized as a complex network in which each supplying location is measured with the EC score to indicate its importance weighted from international customers' point of view. In the second stage, we employ the weighted FN algorithm from modularity-based community structure theory to propose the GCS algorithm, which optimizes hub locations of dry ports and their coverage areas, based on inland regions' factors from the microeconomic (i.e., the EC score rankings), macroeconomic (i.e., foreign trade economics), and geographic (i.e., spatial distance) perspectives. The proposed approach is validated using the real case study of Chinese dry port developments in the context of the BRI. As a result, the optimal locations suggested are closely in line with the real BRI development plans, therefore, the ARMEC-GCS approach is validated.

The contributions of this study are threefold: theoretical, methodological and practical. For the theoretical contribution, many previous studies evaluate the dry port locations based on macroeconomic perspective such as transportation condition, local policy environment and regional economic development, while the assessment based on international customers' perspective is largely overlooked. Furthermore, existings studies focuse mainly on the dry port development at a small scale. Hence, to the best of our knowledge, this paper is the first to explore the location preference from international

customers' perspective and take into account such insights in the decision-making process of large-scale dry port development.

For the methodological contribution, this is a pioneering study applying the data-driven approach for the large-scale dry port location optimization problem. The advantages of our proposed ARMEC-GSC optimization method are as follows:

(1) As compared to classical methods in location theory, such as MCDM and MIP, the novelty of our approach is that the methods used in the ARMEC stage for location importance ranking as well as in the GSC stage for location optimization, are both nonparametric and data-driven without prior assumptions made on the variable distribution. By this way, the location advantage of each inland region can be truly explored in nature by letting the data speak for itself.

(2) Although ARM is a powerful data mining tool to extract hidden patterns out of the large-scale transactional databases, its main drawback is that there may be too many patterns found, which makes the analysis difficult and computationally expensive. Hence, by combining with eigenvector centrality (EC) in the complex network theory, ARM patterns can be visualized as a network of which complex relationships can be analyzed effectively.

(3) The proposed GSC algorithm provides an efficient and realistic optimization approach for the dry port location and allocation problem in the large-scale, complex logistics network. As compared to the classical FN algorithm which was originally developed only for the community structure detection of an unweighted network, we improve it with the gravity function measuring logistics relationships between nodes, so that the proposed CSG algorithm is capable of dealing with a real-world, weighted network.

Regarding the practical contribution, our new effective approach is able to produce a realistic and applicable dry port location solution covering the large-scale area of Mainland China. In particular, the optimal solution is derived from multiple decision-making perspectives (i.e., macroeconomic, microeconomic and geographical), which in turn increases the possibility of its acceptance by various groups of stakeholders and of obtaining funds from the BRI investment, as this solution is practically in line with the market-based principle of the BRI, holding that although the initiative is a policy proposal, its execution must make commercial sense. Furthermore, this paper is expected to help solve the current predicament of the Chinese dry port development, and also serving as a reference guidance for the systematic dry port development in other countries.

This study opens up considerable opportunities to expedite the research progress and the practicability of location theory in the era of Industry 4.0 by adopting new modelling techniques from two emergent domains that have been widely used to study many real-world systems: data mining (also machine learning) and complex network theory. In this regard, a large variety of real-world big data sources

(e.g., Alibaba, Amazon, and eBay) can also be leveraged for new location criteria. Hence, the paper promotes synergies between operation research and data mining – a new, important research stream.

There are some limitations in our research that should be investigated in future research. Firstly, our solution is quite sensitive to different settings of the minimum support and minimum confidence threshold in the ARM model. Hence, future research can improve the model reliability by feeding the optimization component into the ARM model to find optimal values for these parameters. Secondly, our proposed GCS algorithm adopts the hard network divisions for non-overlapping communities, meaning that an inland region can only belong to one community. It would be worthwhile in future studies to investigate dry port networks with overlapping communities, which are also very common in reality.

References

Abbasi, M. and Pishvaee, M.S. (2018), "A two-stage GIS-based optimization model for the dry port location problem : A case study of Iran", *Journal of Industrial and Systems Engineering*.

Anderson, J.E. and van Wincoop, E. (2003), "Gravity with gravitas: a solution to the border puzzle", *American Economic Review*, Vol. 93, pp. 170–192.

Bae, J.K. and Kim, J. (2011), "Product development with data mining techniques: A case on design of digital camera", *Expert Systems with Applications*, Elsevier Ltd, Vol. 38 No. 8, pp. 9274–9280.

Belt & Road News. (2019), "New Freight Train Route links Yantai, Duisburg - Belt & Road News", *Belt & Road News*, available at: https://www.beltandroad.news/2019/07/28/new-freight-train-route-links-yantai-duisburg/ (accessed 14 September 2019).

Belyi, E., Giabbanelli, P.J., Patel, I., Balabhadrapathruni, N.H., Abdallah, A. Ben, Hameed, W. and Mago, V.K. (2016), "Combining association rule mining and network analysis for pharmacosurveillance", *Journal of Supercomputing*, Vol. 72 No. 5, pp. 2014–2034.

Brookes, S. and Huynh, H.N. (2018), "Transport networks and towns in Roman and early medieval England: An application of PageRank to archaeological questions", Journal of Archaeological Science: Reports, Elsevier, Vol. 17 No. December 2017, pp. 477–490.

Campbell, J.F. and O'Kelly, M.E. (2012), "Twenty-Five Years of Hub Location Research", *Transportation Science*, Vol. 46 No. 2, pp. 153–169.

Canh, L. and Notteboom, T. (2016), "A Multi-Criteria Approach to Dry Port Location in Developing Economies with Application to Vietnam", *The Asian Journal of Shipping and Logistics*.

Chang, Z., Notteboom, T. and Lu, J. (2015), "A two-phase model for dry port location with an application to the port of Dalian in China", *Transportation Planning and Technology*, Vol. 38 No. 4, pp. 442–464.

Chen, H., Lam, J.S.L. and Liu, N. (2018), "Strategic investment in enhancing port–hinterland container transportation network resilience: A network game theory approach", *Transportation Research Part B: Methodological*, Vol. 111, pp. 83–112.

Chiang, D.M.-H., Lin, C.-P. and Chen, M.-C. (2011), "The adaptive approach for storage assignment by mining data of warehouse management system for distribution centres", *Enterprise Information Systems*, Vol. 5 No. 2, pp. 219–234.

China Daily. (2015), "Guangdong sees big role in 'One Belt, One Road' - China - Chinadaily.com.cn", *China Daily*, available at: http://www.chinadaily.com.cn/regional/2015-

06/05/content_20930718.htm (accessed 15 September 2019).

China Daily. (2017), "Yantai awarded as one of most dynamic Belt and Road cities", *China Daily*, available at: http://www.chinadaily.com.cn/m/shandong/yantai/2017-01/16/content_27962935.htm (accessed 14 September 2019).

China Daily. (2018), "Guangdong – A key hub on the Maritime Silk Road - Opinion - Chinadaily.com.cn", *China Daily*, available at: http://www.chinadaily.com.cn/a/201809/25/WS5ba9e159a310c4cc775e7fbc.html (accessed 15 September 2019).

Clauset, A., Newman, M.E.J. and Moore, C. (2004), "Finding community structure in very large networks", *Physical Review E*, Vol. 70 No. 6, p. 6.

Cohen, M.C. (2018), "Big Data and Service Operations", *Production and Operations Management*, Vol. 27 No. 9, pp. 1709–1723.

Corne, D., Dhaenens, C. and Jourdan, L. (2012), "Synergies between operations research and data mining: The emerging use of multi-objective approaches", *European Journal of Operational Research*, Vol. 221 No. 3, pp. 469–479.

Costa, A. (2015), "MILP formulations for the modularity density maximization problem", *European Journal of Operational Research*, Elsevier Ltd., Vol. 245 No. 1, pp. 14–21.

Crainic, T.G., Dell'Olmo, P., Ricciardi, N. and Sgalambro, A. (2015), "Modeling dry-port-based freight distribution planning", *Transportation Research Part C: Emerging Technologies*, Vol. 55, pp. 518–534.

DHL. (2016), *"Belt and Road": What You Need to Know*, available at: https://www.logistics.dhl/content/dam/dhl/global/dhl-global-forwarding/documents/pdf/dhl-glo-dgf-belt-and-road.pdf (accessed 14 September 2019).

El-adaway, I.H., Abotaleb, I. and Vechan, E. (2018), "Identifying the most critical transportation intersections using social network analysis", Transportation Planning and Technology, Taylor & Francis, Vol. 41 No. 4, pp. 353–374.

en.people.cn. (2018), "The Belt and Road gives boost to Ningbo-Zhoushan port - People's Daily Online", *En.People.Cn*, available at: http://en.people.cn/n3/2018/0814/c90000-9490588.html (accessed 14 September 2019).

Ettl, M., Harsha, P., Papush, A. and Perakis, G. (2019), "A Data-Driven Approach to Personalized Bundle Pricing and Recommendation", *Manufacturing & Service Operations Management*.

Feng, X., Zhang, Y., Li, Y. and Wang, W. (2013), "A location-allocation model for seaport-dry port system optimization", *Discrete Dynamics in Nature and Society*.

Ghanbari, R., Jalili, M. and Yu, X. (2018), "Correlation of cascade failures and centrality measures in complex networks", *Future Generation Computer Systems*, Vol. 83, pp. 390–400.

Ghedini Ralha, C. and Sarmento Silva, C.V. (2012), "A multi-agent data mining system for cartel detection in Brazilian government procurement", *Expert Systems with Applications*, Vol. 39 No. 14, pp. 11642–11656.

Govindan, K., Cheng, T.C.E., Mishra, N. and Shukla, N. (2018), "Big data analytics and application for logistics and supply chain management", *Transportation Research Part E: Logistics and Transportation Review*, Vol. 114, pp. 343–349.

HKTDC research. (2019), "Guangdong: Market Profile", *HKTDC Research*, available at: http://china-trade-research.hktdc.com/business-news/article/Facts-and-Figures/Guangdong-Market-Profile/ff/en/1/1X000000/1X06BUOU.htm (accessed 15 September 2019).

Huang, Y. (2016), "Understanding China's Belt & Road Initiative: Motivation, framework and assessment", *China Economic Review*, Vol. 40, pp. 314–321.

Ishfaq, R. and Sox, C.R. (2011), "Hub location-allocation in intermodal logistic networks", *European Journal of Operational Research*, Vol. 210 No. 2, pp. 213–230.

Ka, B. (2011), "Application of fuzzy AHP and ELECTRE to China dry port location selection", *Asian Journal of Shipping and Logistics*, Vol. 27 No. 2, pp. 331–354.

Kaluza, P., Kölzsch, A., Gastner, M.T. and Blasius, B. (2010), "The complex network of global cargo ship movements", *Journal of the Royal Society Interface*, pp. 1093–1103.

khosravi, S. and Akbari Jokar, M.R. (2017), "Facility and hub location model based on gravity rule", *Computers and Industrial Engineering*, Elsevier Ltd, Vol. 109, pp. 28–38.

Komchornrit, K. (2017), "The Selection of Dry Port Location by a Hybrid CFA-MACBETH-PROMETHEE Method : A Case Study of Southern Thailand", *The Asian Journal of Shipping and Logistics*, Vol. 33 No. 3, pp. 141–153.

Konings, J. (2018), *Trade Impacts of the Belt and Road Initiative*, available at: www.ing.com/THINK (accessed 29 March 2019).

KPMG China. (2018), "A New Xi'an in the New Era — KPMG Contributes to the Development of a New Reform Locomotive in Inland China - KPMG China", *KPMG China*, 29 June, available at: https://home.kpmg/cn/en/home/news-media/press-releases/2018/06/a-new-xi-an-in-the-new-era.html (accessed 13 September 2019).

De La Iglesia, B., Richards, G., Philpott, M.S. and Rayward-Smith, V.J. (2006), "The application and effectiveness of a multi-objective metaheuristic algorithm for partial classification", *European Journal of Operational Research*, Vol. 169 No. 3, pp. 898–917.

Lättilä, L., Henttu, V. and Hilmola, O. (2015), "Hinterland operations of sea ports do matter : Dry port usage effects on transportation costs and CO 2 emissions", *Transportation Research Part E: Logistics and Transportation Review*, Vol. 55 No. 2013, pp. 23–42.

Lee, C.K.H. (2016), "A GA-based optimisation model for big data analytics supporting anticipatory shipping in Retail 4.0", *International Journal of Production Research*, Vol. 7543, pp. 1–13.

Lee, C.Y. and Song, D.P. (2017), "Ocean container transport in global supply chains: Overview and research opportunities", *Transportation Research Part B: Methodological*, Elsevier Ltd, Vol. 95, pp. 442–474.

Lehman Brown International Accountants. (2017), *The Belt and Road Initiative*, *Lehman Brown International Accountants*, available at:https://doi.org/10.17265/2160-6579/2016.02.002.

Li, F., Shi, X. and Hu, H. (2011), "Location selection of dry port based on AP clustering: The case of SouthWest China", *Journal of System and Management Sciences*, Vol. 2 No. 5, pp. 255–261.

Liu, W., Shen, X. and Wang, D. (2018), "The impacts of dual overconfidence behavior and demand updating on the decisions of port service supply chain: a real case study from China", *Annals of Operations Research*.

Liu, Y., Liu, G., Liu, Q. and Qin, Z. (2013), "Community Detection in Real Large Directed Weighted Networks", *International Journal of Digital Content Technology and Its Applications*, Vol. 7 No. 5, pp. 521–529.

Misic, V. and Perakis, G. (2019), "Data Analytics in Operations Management: A Review", *Manufacturing & Service Operations Management*.

Newman, M.E.J. (2004a), "Analysis of weighted networks", *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, Vol. 70 No. 5, p. 9.

Newman, M.E.J. (2004b), “Fast algorithm for detecting community structure in networks”, *Physical Review E*, Vol. 69 No. 6, p. 5.

Newman, M.E.J. and Girvan, M. (2004), “Finding and evaluating community structure in networks”, *Physical Review E*, available at:https://doi.org/10.1103/PhysRevE.69.026113.

Ng, K.Y.A. and Gujar, G.C. (2009), “The spatial characteristics of inland transport hubs: evidences from Southern India”, *Journal of Transport Geography*, Vol. 17 No. 5, pp. 346–356.

Nguyen, T. Van, Zhou, L., Chong, A.Y.L., Li, B. and Pu, X. (2019), “Predicting customer demand for remanufactured products: A data-mining approach”, *European Journal of Operational Research*.

Nguyen, T., ZHOU, L., Spiegler, V., Ieromonachou, P. and Lin, Y. (2018), “Big data analytics in supply chain management: A state-of-the-art literature review”, *Computers & Operations Research*, Pergamon, Vol. 98, pp. 254–264.

Padmanabhan, B. and Tuzhilin, A. (2003), “On the Use of Optimization for Data Mining: Theoretical Interactions and eCRM Opportunities”, *Management Science*, Vol. 49 No. 10, pp. 1327–1343.

Parajuli, J. and Haynes, K.E. (2018), “Transportation network analysis in Nepal: a step toward critical infrastructure protection”, Journal of Transportation Security, Journal of Transportation Security, Vol. 11 No. 3–4, pp. 101–116.

Post and Parcel. (2016), “DHL supporting Chengdu as part of China’s "Belt and Road" initiative | Post & Parcel”, *Post and Parcel*, available at: https://postandparcel.info/73296/news/dhl-supporting-chengdu-as-part-of-chinas-belt-and-road-initiative/ (accessed 15 September 2019).

Qiu, X. and Lee, C.-Y. (2019), “Quantity discount pricing for rail transport in a dry port system”, *Transportation Research Part E: Logistics and Transportation Review*, Pergamon, Vol. 122, pp. 563–580.

Railwaypro.com. (2017), “Rail transport is cheaper than air but faster than sea”, *Railwaypro.Com*, available at: https://www.railwaypro.com/wp/rail-transport-cheaper-air-faster-sea/ (accessed 13 September 2019).

Roso, V. and Lumsden, K. (2010), “A review of dry ports”, *Maritime Economics and Logistics*, Vol. 12 No. 2, pp. 196–213.

Roso, V., Woxenius, J. and Lumsden, K. (2009), “The dry port concept: connecting container seaports with the hinterland”, *Journal of Transport Geography Journal*, Vol. 17, pp. 338–345.

Rubinov, M. and Sporns, O. (2010), “Complex network measures of brain connectivity: Uses and interpretations”, *NeuroImage*, Vol. 52 No. 3, pp. 1059–1069.

Ruiz, R., Asgari, N., Farahani, R.Z., Fallah, S. and Hosseini, S. (2018), “OR models in urban service facility location: A critical review of applications and future developments”, *European Journal of Operational Research*, Elsevier B.V., Vol. 276 No. 1, pp. 1–27.

Saberi, M., Mahmassani, H.S., Brockmann, D. and Hosseini, A. (2017), “A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large-scale origin–destination demand networks”, *Transportation*, Vol. 44 No. 6, pp. 1383–1402.

Samaa Digital. (2017), “‘One Belt, One Road’ brings new opportunities for former Silk Road city of Jiayuguan - Samaa Digital”, *Samaa Digital*, available at: https://www.samaa.tv/economy/2017/07/one-belt-one-road-brings-new-opportunities-for-former-silk-road-city-of-jiayuguan/ (accessed 14 September 2019).

Santiago, R. and Lamb, L.C. (2017), “Efficient modularity density heuristics for large graphs”, *European Journal of Operational Research*, Vol. 258 No. 3, pp. 844–865.

Sarker, M.N.I., Hossin, M.A., Yin, X. and Sarkar, M.K. (2018), "One Belt One Road Initiative of China: Implication for Future of Global Development", *Modern Economy*, Vol. 09 No. 04, pp. 623–638.

Sarwar, F. (2018), "China's One Belt and One Road: Implications of 'New Eurasian Land Bridge' on Global Power Play in the Region", *NUST Journal of International Peace & Stability*, Vol. 1 No. 2, pp. 131–144.

Silin, Y., Kapustina, L., Trevisan, I. and Drevalev, A. (2017), "China's economic interests in the 'One Belt, One Road' initiative", *SHS Web of Conferences*, Vol. 39, p. 01025.

Sun, Z., Zheng, J. and Hu, H. (2012), "Finding community structure in spatial maritime shipping networks", *International Journal of Modern Physics C*, Vol. 23 No. 06.

Swiss Re Institute. (2017), *China's Belt & Road Initiative: The Impact on Commercial Insurance in Participating Regions*.

The Ministry of Transport of the People's Republic of China. (2017), *Development of International Dry Port in China*, available at: https://www.unescap.org/sites/default/files/China_EGM Dry Ports_2017.pdf.

Ting, S.L., Tse, Y.K., Ho, G.T.S., Chung, S.H. and Pang, G. (2014), "Mining logistics data to assure the quality in a sustainable food supply chain: A case in the red wine industry", *International Journal of Production Economics*, Vol. 152, pp. 200–209.

Tsai, C.-Y. and Huang, S.-H. (2015), "A data mining approach to optimise shelf space allocation in consideration of customer purchase and moving behaviours", *International Journal of Production Research*, Vol. 53 No. 3, p. 850.

Tsao, Y.C. and Thanh, V. Van. (2019), "A multi-objective mixed robust possibilistic flexible programming approach for sustainable seaport-dry port network design under an uncertain environment", *Transportation Research Part E: Logistics and Transportation Review*, Elsevier, Vol. 124, pp. 13–39.

Verma, P., Nandi, A.K. and Sengupta, S. (2018), "Bribery games on interdependent complex networks", *Journal of Theoretical Biology*, Vol. 450, pp. 43–52.

Wang, C., Chen, Q. and Huang, R. (2018), "Locating dry ports on a network: a case study on Tianjin Port", *Maritime Policy and Management*, Vol. 45 No. 1, pp. 71–88.

Wang, W., Li, Z. and Cheng, X. (2018), "Evolution of the global coal trade network: A complex network analysis", *Resources Policy*, No. 1.

Wang, X. and Meng, Q. (2019), "Optimal price decisions for joint ventures between port operators and shipping lines under the congestion effect", *European Journal of Operational Research*, Elsevier B.V., Vol. 273 No. 2, pp. 695–707.

Wei, H. and Sheng, Z. (2017), "Dry ports-seaports sustainable logistics network optimization: Considering the environment constraints and the concession cooperation relationships", *Polish Maritime Research*, Vol. 24 No. S3, pp. 143–151.

Wei, H., Sheng, Z. and Lee, P.T.W. (2018), "The role of dry port in hub-and-spoke network under Belt and Road Initiative", *Maritime Policy and Management*, Vol. 45 No. 3, pp. 370–387.

Witte, P., Wiegmans, B. and Ng, A.K.Y. (2019), "A critical review on the evolution and development of inland port research", *Journal of Transport Geography*, Elsevier, Vol. 74, pp. 53–61.

Witten, I.H. and Frank, E. (2011), *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, available at:https://doi.org/0120884070, 9780120884070.

Xie, Y., Liang, X., Ma, L. and Yan, H. (2017), "Empty container management and coordination in intermodal transport", *European Journal of Operational Research*, Elsevier B.V., Vol. 257 No. 1,

pp. 223–232.

Yang, D., Pan, K. and Wang, S. (2017), “On service network improvement for shipping liners shipping lines under the one belt one road initiative of China”, *Transportation Research Part E: Logistics and Transportation Review*, pp. 1–14.

Yu, M., Fransoo, J.C. and Lee, C.Y. (2018), “Detention decisions for empty containers in the hinterland transportation system”, *Transportation Research Part B: Methodological*, Vol. 110, pp. 188–208.

Yu, W., Chavez, R., Jacobs, M.A. and Feng, M. (2018), “Data-driven supply chain capabilities and performance: A resource-based view”, *Transportation Research Part E: Logistics and Transportation Review*, Elsevier Ltd, Vol. 114, pp. 371–385.

Zeng, Q., Wang, G.W.Y., Qu, C. and Li, K.X. (2017), “Impact of the Carat Canal on the evolution of hub ports under China’s Belt and Road initiative”, *Transportation Research Part E: Logistics and Transportation Review*, available at:https://doi.org/10.1016/j.tre.2017.05.009.

Zhang, J. and Meng, M. (2019), “Bike allocation strategies in a competitive dockless bike sharing market”, *Journal of Cleaner Production*, Elsevier Ltd, Vol. 233, pp. 869–879.

Zhang, Q., Wang, W., Peng, Y., Zhang, J. and Guo, Z. (2018), “A game-theoretical model of port competition on intermodal network and pricing strategy”, *Transportation Research Part E: Logistics and Transportation Review*, Elsevier, Vol. 114, pp. 19–39.

Zheng, J., Qi, J., Sun, Z. and Li, F. (2018), “Community structure based global hub location problem in liner shipping”, *Transportation Research Part E: Logistics and Transportation Review*, Vol. 118, pp. 1–19.

Zhou, H., Zhang, Y. and Li, J. (2018), “An overlapping community detection algorithm in complex networks based on information theory”, *Data & Knowledge Engineering*.

Zhu, Y., Zhou, L., Xie, C., Wang, G.-J. and Nguyen, T. V. (2019), “Forecasting SMEs’ credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach”, *International Journal of Production Economics*, Elsevier, Vol. 211, pp. 22–33.