Using k-means clustering in international location decision

Khalid, Waqas; Herbert-Hansen, Zaza Nadja Lee

Published in:
Journal of Global Operations and Strategic Sourcing

Link to article, DOI:
10.1108/JGOSS-11-2017-0056

Publication date:
2018

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Using k-means clustering in international location decision

Published in: Journal of Global Operations and Strategic Sourcing, April 2018
Waqas Khalid & Zaza Nadja Lee Hansen

Abstract

Purpose This article investigates the application of unsupervised machine learning in the international location decision (ILD). This paper addresses the need for a fast, quantitative, and dynamic location decision framework.

Methodology/Approach Unsupervised machine learning technique i.e. k-means clustering is used to carry out the analysis. 24 different indicators of 94 countries, categorized into five groups, have been used in the analysis. After the clustering, the clusters have been compared and scored to select the feasible countries.

Findings A new framework is developed based on k-means clustering that can be used in ILD. This method provides a quantitative output without personal subjectivity. The indicators can be easily added or extracted based on the preferences of the decision makers. Hence, it was found out that the unsupervised machine learning i.e. k-means clustering is a fast and flexible decision support framework that can be used in ILD.

Limitations Limitations include the generality of selected indicators and clustering algorithm used. The use of other methods and parameters may lead to alternate results.

Originality/Value The framework developed through the research intends to assist the decision makers in deciding on the facility locations. The framework can be used in international as well national domains. It provides a quantitative, fast, and flexible way to shortlist the potential locations. Other methods can also be used to further decide on the specific location.

Keywords International location decision, K-mean, Clustering, Machine learning, Facility location, Supply chain, Offshoring

Paper type Research paper

Introduction

Companies today are competing in global markets, which brings extra pressure and competition, forcing companies to consider international location decision (De Toni et al., 1992, Maccarthy and Atthirawong, 2015, Meijboom and Vos, 1997). An international location decision (ILD) is a macro-level decision made by a firm, to pinpoint the geographic location of a partly or fully owned facility from a set of international alternatives (Schmidt et al., 2017). International alternatives can also be feasible locations within one country. Choosing the right strategic location across the globe may result in reducing operational costs and maximizing closeness to customers, hence the importance of strategic location cannot be overlooked (Behar and Venable, 2011, Bogataj et al., 2011, Ellram et al., 2013, Fishman, 2012). The main motives behind offshoring manufacturing facilities are to take advantage of economies of scale or entry into new markets (Badri, 1999, Sinha et al., 2011). Whereas, other factors are access to low-cost labor, material, and components to maximize profit; access to local technological resources; and proximity to markets for better customer service (Dunning, 1994, Ferdows, 1997, Maccarthy and Atthirawong, 2015, Canham and Hamilton, 2013). Moreover, today international facility location is not limited to only manufacturing or warehouse facilities, but other service factors like R&D facility, call centers, ITeS, or BPO etc. are also considered (Viswanadham and Kameshwaran, 2007). This implies international facility location is company specific and variation in influencing factors can be large.

Considering the large variation in the influencing factors, there is a need of a model or technique that is flexible enough to adapt to changing requirements specific to each company. Several tools and techniques exist in the global supply chain that companies executives can use in making an international location decision. For example, qualitative techniques such as PERC Model (Viswanadham and Kameshwaran, 2007), Delphi Technique (Maccarthy and Atthirawong, 2015), The Scope Model (Schmidt et al., 2017), and Roadmapping & Mountain Model (Christodoulou et al., 2007). These techniques are either based on direct or indirect observations or experiences and provide a qualitative output or are visualization or planning frameworks. Such techniques are also prone to personal subjectivity. Several numerical or analytical techniques also exist that are used in facility location decision and provide a quantitative output, such as, ROI (Biswas et al., 2016) and AHP (Viswanadham and Kameshwaran, 2007). Many of these techniques are optimization techniques, such as Goal Programming (Badri, 1999), Fuzzy Group Assessment (Lin and Lee, 2007), and Stochastic Model (Goh et al., 2007). These techniques are either complex enough that they do not allow large flexibility in the influencing factors or they require prior shortlisting of few alternatives that are to be considered in decision-making process. There are many cost optimization techniques as well such as Mixed Integer Programming Technique (Cui et al., 2010), Baumol-Wolfe Method, Kuehn-Hamburger Model, Gravity Method, and CFLP Method (Liu et al., 2010). These optimization
techniques mainly focus on optimizing transportation costs and do not consider macro logistics factors such as economic, political, or infrastructural aspects. Some authors have used risk aspects in facility locations such as Grey based MCDM (Prakash et al., 2015) and Stochastic Model (Goh et al., 2007). These model either prone to personal subjectivity or too complex to integrate a large number of influencing variables. Hence, there is a need for an alternate solution that can allow company's executives to consider the flexible and large number of factors in ILD. Further, the method can generate quantitative output without personal subjectivity and prior need of shortlisting of few alternatives.

Considering the ability of machine learning to handle large and flexible data without personal subjectivity can be an answer to this problem. Machine learning is rapidly making its way into every field of life (Bobrow and Brady, 1999, Cheng et al., 2016, Fox, 1986, Schorr and Rappaport, 1989, Shahrabi et al., 2009, Wang et al., 2017, Zage et al., 2013). However, researchers have limited themselves only to a few fields of the supply chain. Most of the applications are in the area of forecasting, supply chain risk, and vendor selection (Li, 2012, Shahrabi et al., 2009, Zage et al., 2013). However, several attempts have been made to extend machine learning applications in global supply chain sector (Jayaram, 2016, Witkowski, 2017), but scholars have paid far too little attention to the field of International Facility Location. Availability of location data, for example, countries economic or infrastructure indicators play a significant role in making facility location decision (Maccarthy and Atthirawong, 2015). Moreover, at the same instance availability of the data is the core strength of machine learning as well (Tzanis et al., 2008). Thus, this fact leads us to question on how this available data can be used in international location decision process using machine learning technique. And this brings us to the research question of this article: while only focusing on the unsupervised learning:

Research question (RQ): “How can unsupervised machine learning be used in international location decision (ILD) process?”

Note that this paper focuses only on unsupervised learning because the nature of the problem at hand does not allow for a supervised learning analysis as the goal is not to predict the location but to group the locations with similar attributes and then decide based on the results.

This article is structured as follows; first, a literature review details the latest findings within ILD. The machine learning technique used to carry out analysis is explained in the methodology section. The data collected and its limitations have been described. A thorough analysis of the data is carried out and results are presented in the findings. Reflection is made on limitations and a comprehensive conclusion is presented at the end.

Literature Review

In this article, two main literature review strategies are used. The first is the pearl growing technique, where the most relevant article was found and then used to find the other related sources. The second technique is the bibliographic mining, where reviewing of the references of relevant articles was carried out to mine the relevant sources. Table 1 shows some of the facility location frameworks developed in last two decades along with their strengths and weaknesses.

Table 1. Literature review of models and frameworks

In the global supply chain, International manufacturing or offshoring manufacturing facilities has become a central issue for executives of companies (Badri, 1999, Farrell, 2006). It is becoming increasingly difficult for companies to ignore the importance of International Location Decision (Dunning, 1994, Ferdows, 1997, Maccarthy and Atthirawong, 2015). Ferdows (1997) highlights the importance of establishing foreign factories by classifying them into several categories based on strategic reasons for the site. Poorly placed activity cells (production, distribution, and reverse logistic plants) in a global supply chain network can result in excessive direct transportation and other related production costs no matter how good inventory policies, transportation plans, and information sharing policies are. Mixed integer based extended MRP model can be used to study the effect of location of these cells in the global supply chain (Bogataj et al., 2011, Cui et al., 2010, Liu et al., 2010). International Location Decision is one of the most researched areas in the field of the global supply chain (Lu et al., 2014, Maccarthy and Atthirawong, 2015, Tucker and Armstrong, 2008, Zhang et al., 2010). Several numerical and empirical methods exist that are used in ILD process. For example, PERC, a qualitative model (Viswanadham and Kameshwaran, 2007). They used analytical hierarchy process (AHP), which is a quantitative model, to cluster best possible alternatives under four fundamental criteria: product/process value chain, economic and political integration, resources and management, and connecting technologies. Lin and Lee (2007) have proposed an
unbiased fuzzy group assessment model that decision-makers can use to decide on the location, based on quantitative output, without personal subjectivity. Several linear programming models and heuristics have been proposed by many researchers in facility location decision. Most of them are based on optimization problem based on single or multiple objectives. For example, Badri (1999) has proposed AHP and GP (Goal Programming) approach that aid in decision makers to select an appropriate facility location by weighing in and optimizing multiple criteria. Mixed Integer Programming (MIP) techniques can be used to optimize facility location so that the setup and transportation costs are minimized and customers are assigned optimally (Cui et al., 2010). Liu et al. (2010) have mentioned several optimization methods such as Baumol-Wolf, Kuehn-Hamburger, Gravity, and CFLP methods to optimally select facility cells in supply chain network. These methods mainly focus on micro logistics and transportation-related costs and do not consider other economic, political, or infrastructure aspects at macro logistics level.

Maccarthy and Athirawong (2015) have highlighted several critical factors involved in an ILD process based on qualitative Delphi Approach including several experts and consultants in the field. Some of the main factors are favourable labor climate, transportation costs, proximity to market and customers, and proximity to suppliers and resources etc. Web-based decision support system can also be used for plant location decision. In this model, all relevant data of plant locations are stored in server database by industrial parks or individual companies and when required, the data is extracted by decision makers and based on requirements a set of feasible location is selected (Weng et al., 2011). M. Telsang proposed a quantitative technique called dimensional analysis where costs for two alternatives are calculated by weighting factors. Relative Merit (RM) is then calculated by a fraction of these two costs and the alternative with less cost is selected (Biswas et al., 2016). ROI (return on investment) is another quantitative technique where alternatives are evaluated and the location is selected with the highest ROI value (Biswas et al., 2016). Factor Point Matrix technique designed by Merill R. can be used in the location decision. In this technique, weightage is given to all factors carefully and individual scores are given to each alternative against each factor. Scores are multiplied with weight and summed together and the alternative with the highest score is selected (Biswas et al., 2016).

Offshoring also brings many challenges and risks (Sharma and Kaushik, 2017). Prakash et al. (2015) have proposed a different and unique framework for facility location decision. They have proposed a Grey based Multi-Criteria Decision Making (MCDM) model to evaluate a set of alternatives based on the risks and uncertainties. A set of alternatives, based on company's supply chain strategy, are selected. Risk enabling factors (REF) are identified and each alternative is evaluated against each REF. The alternatives are then ranked and the one with the least risk is selected. Goh et al. (2007) have also considered risks factors in supply chain networks. The stochastic model proposed by Goh et al. (2007) uses operational research techniques to minimize risks and maximize profits. But, this model does not consider other economic, infrastructure, or political factors. Moreover, the model can be very large and complex for higher numbers of influencing factors.

A scope model has been proposed by Schmidt et al. (2017). This model proposes a comprehensive approach towards ILD that can be used by decision makers. The model focuses on 4 aspects i.e. State, Organizational, Change and Alternatives aspects. This model states that the abstraction level increases while moving closer to the center of the model. Hence, making it easier for the decision makers to deciding on facility location. Roadmapping is another model that is useful in visualizing and planning. The firms can use this model in aligning the business strategy and manufacturing footprints. However, this model does not provide any quantitative output (Christodoulou et al., 2007). Mountain Model is another framework that can be helpful in deciding on the roles of the plant in supply chain network (Christodoulou et al., 2007). Windmark and Andersson (2016) have proposed a cost model that compares the total cost to produce one part for each alternative. The alternative with the least cost is selected.

Applications of machine learning in supply chain

New applications are emerging in the field of AI (Artificial Intelligence) and Machine Learning, which are revolutionizing common lives. The supply chain is one the most significant fields of research in machine learning. Shahrabi et al. (2009) have highlighted the importance of machine learning in supply chain forecasting by comparing sales forecasting efficiency of conventional methods and more advanced machine learning algorithms like Artificial Neural Network (ANN) and Support Vector Machine (SVM). Carbonneau et al. (2008) have analyzed the effect of distorted demand at the end of the supply chain (bullwhip effect) and found out that machine learning algorithms SVM and RNN (Recurrent Neural Networks) outperformed other methods. Some efforts are also been made in Supply chain risk management using machine learning mathematical models that may detect deception or deceptive practices and can be integrated into a company's supply chain network (Zage et al., 2013). Li (2012) has proposed a mathematical model based on Support Vector Machine (SVM) to strategically select supply chain partners. This approach minimizes the human interference and optimizes the efficiency in partner evaluation.
Although the research has been made in the several fields of the supply chain sector, no attempt has been made to apply machine learning techniques in international location decision process. Hence, this article will focus on the application of machine learning algorithms in facility location process.

**Methodology**

In this article, it has been hypothesized that k-mean clustering can be used as a decision framework in ILD. The k-means clustering has been used to group countries with similar attributes into clusters. Clustering is an unsupervised machine learning technique (Haroon, 2017, Louridas and Ebert, 2016, Sotiropoulos and Tsihrintzis, 2017), which means, unlike supervised machine learning technique it is not allowed to validate predicted output due to unlabelled data (Ozgur, 2004). Many clustering techniques are used in practice such as k-means, DBSCAN, Agglomerative, Divisive (DIANA), or COBWEB (Mehta et al., 2015). In this article, k-means clustering is used to cluster countries based on the similarities among the influencing factors. The methodology used, is inspired by Ng (2017) with few modifications that is explained below:

**Step 1:** Countries data is collected based on the economic, infrastructural, political, and market indicators. \( x_{(j)}^{(i)} \) represents the value of \( j \)th indicator for the \( i \)th country.

\[
x_{(j)}^{(i)} = \text{Value of } j\text{th indicator of } i\text{th country}
\]

\[
\forall i = 1,2,3 \ldots m, j = 1,2,3 \ldots n
\]

\[
m = \text{Number of countries}
\]

\[
n = \text{Number of indicators}
\]

**Step 2:** Indicators may have different units and values may differ significantly to one another. Hence, it is important to normalize or feature scale the data so that all countries may have values that are in comparable range to each other. Mean normalization technique is used in this case. Mean \( \mu(j) \) is calculated for each indicator \( j \) by using Equation 1.

\[
\mu(j) = \frac{1}{m} \sum_{i=1}^{m} x_{(j)}^{(i)} \tag{1}
\]

Then each value \( x_{(j)}^{(i)} \) is replaced by \( x_{(j)}^{(i)} - \mu(j) \).

**Step 3:** Now that the data has been normalized, in the next step Principal Component Analysis (PCA) is used to reduce the dimensionality of the data. As the data may have tens to hundreds of indicators as features \( (n) \), on the basis of which the countries \( (m) \) have to be clustered. Hence, to get rid of unnecessary features (indicators \( n \)) and reduce the dimensionality to a lower value of \( d \), PCA was used for the sake of simplicity and visibility. Whereas, \( d \) is the reduced number of dimensions. In PCA, data is projected onto \( d \) “eigenvectors” \( u^{(1)}, u^{(2)}, \ldots, u^{(d)} \) so that the projection error is minimum. In order to reduce the dimensionality, a covariance matrix “Sigma” is calculated by Equation 2.

\[
Sigma = \Sigma = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)}) \cdot (x^{(i)})^{T} \tag{2}
\]

To compute the “eigenvectors” of matrix “Sigma”, MATLAB function "[U, S, V] = svd(Sigma)" is used. This function takes the covariance vector “Sigma” as input and returns \( U \) as a set of “eigenvectors” \( u^{(1)}, u^{(2)}, \ldots, u^{(n)} \) for \( n \) features. These "eigenvectors" are the unit vectors that remain unchanged when the scalar values of the vectors are transformed. The data is then projected onto these "eigenvectors" to get the lower dimensionalized data.

**Step 4:** In order to reduce the data from \( n \) to \( d \) dimensions, \( d \) eigenvectors are to be selected. For this purpose, MATLAB function \( U_{reduce} = U(:, 1:d) \) is used. Then \( U_{reduce} \) is multiplied with matrix \( X \) to project the countries data onto \( d \) eigenvectors by using Equation 3.

\[
Z = X \cdot U_{reduce} \tag{3}
\]
\[ U_{\text{reduce}} = \text{Reduced eigenvector matrix of dimension } n \times d \]
\[ X = \text{Normalized matrix of countries data with dimension } m \times n \]
\[ Z = \text{Reduced matrix of dimension } m \times d \]

Z is a reduced matrix of dimension \( m \times d \) representing a new number of features \( d \) instead of original \( n \) features or indicators.

Step 5: Now country data has been reduced from dimension \( n \) to \( d \) and next step is to check the efficiency of this reduced data. One method is to check the degree of variance retained. Equations 4 is used to check the variance retention.

\[
\frac{\frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - x_{\text{app}}^{(i)}||^2}{\frac{1}{m} \sum_{i=1}^{m} ||x^{(i)}||^2} = Z
\]

(4)

The numerator represents averaged squared projected error and the denominator represents total variation in the data. Whereas, \( x_{\text{app}}^{(i)} \) is the retained value of \( x^{(i)} \) that was reconstructed from reduced data \( Z^{(i)} \) and is calculated by Equation 5.

\[
X_{\text{app}} = Z \cdot (U_{\text{reduce}})^T
\]

(5)

\( X_{\text{app}} = \text{Reconstructed matrix of } x_{\text{app}} \text{ values of dimension } m \times n \)

If the value of equation 4 is less than equal to 0.05 then it shows that 95% of the variance is retained. In this article, minimum of 95% variance retention is used, but this may differ based on the personal subjectivity. In this article, MATLAB function \([U,S,V] = \text{svd}(\text{Sigma})\) is used to calculate the variance retention instead of Equation 4. This function returns a matrix \( S \) of the same dimension as matrix \( \text{Sigma} \) with nonnegative diagonal values and can be used to calculate variance retention by using Equation 6.

\[
\frac{\sum_{i=1}^{d} S_{(i)}}{\sum_{i=1}^{n} S_{(i)}} \quad \forall \ i = 1,2,3\ldots n
\]

(6)

Step 6: If variance retention is less than 95% as calculated from equation 6 then the number of dimension \( d \) is increased and step 4 and 5 are repeated until at least 95% variance retention is achieved.

Step 7: The number of clusters, in which the data has to be divided, is decided. Several methods exist that can be used in deciding the number of clusters \( k \). In this article, two methods are used called Elbow Method (Bholowalia and Kumar, 2014) and Calinski-Harabasz Method (Calinski and Harabasz, 1974). In Elbow method, the sum of squared errors (SSE) is plotted against the number of clusters and the number of clusters \( k \) is selected so that adding more clusters will not provide better results. Calinski-Harabasz method takes the input data and outputs the optimal number of clusters \( k \).

Step 8: In the next step, in order to run k-means clustering algorithm, \( k \) cluster centroids \( \mu_1, \mu_2, \ldots \mu_k \) are initialized. Whereas, \( k \) represents the total number of clusters. In order to avoid, ending up in local minima, initial centroids were randomly picked from the data set. In MATLAB the function \( \text{randidx} = \text{randperm(size}(Z,1)) \) was used to randomize the indices of \( Z \) i.e. reduced feature matrix. And then the function \( \text{centroids} = Z(\text{randidx}(1:K),:) \) was used to pick the first \( k \) indices as cluster centroids.
Step 9: In the next step, the distances between centroids $\mu_1, \mu_2, \ldots, \mu_k$ and individual values $z^{(i)}$ of matrix $Z$ i.e. reduced feature matrix are calculated using Equation 7. And then the index $c^{(i)}$ of centroid $\mu_k$ is assigned to example $z^{(i)}$ for which the distance is the closest. Whereas, $c^{(i)} = 1, 2, \ldots k$

$$\left| z^{(i)} - \mu_k \right|^2$$

\forall i = 1, 2, 3, \ldots m & k = 1, 2 \ldots k

Step 10: In the next step, new centroids are calculated by taking the average of all the data points $z^{(i)}$ which are assigned the same index value $c^{(i)}$. And these mean values are assigned as new centroids.

Step 11: Step 8 and 9 are repeated until more iterations do not result in the change of centroids. These centroids can be termed as satisfactory centroids.

Step 12: The number of clusters assumed in step 7 is validated. In this article, Silhouette plots are used to validate the number of clusters $k$.

Step 13: Clustered data is analyzed and compared. Scores are assigned to each cluster and the one with the highest scores is selected as a set of potential facility location. Figure 1 represents the summary of all the steps detailed above.

**Figure 1. Summary of methodology**

**Data**

Several authors have proposed some key factors to be considered while making facility location decision. For example, Viswanadham and Kameshwaran (2007) have proposed some fundamental criteria consisting of value chain, economic and political integration, resource and management, and connecting technologies. These factors may include labor force, economic or political stability, air or network connectivity etc. Maccarthy and Atthirawong (2015) have also proposed some critical factors in ILD using Delphi study consisting of experts and consultants. Some of the factors proposed are the quality of labor, proximity to market, supplier quality, quality of life, and tax structures etc. In this article, a total of 24 indicators are used. These indicators are inspired by the previous studies. Being a large number of indicators, these have been grouped into sub-categories. These categories consist of economic indicators, political indicators, infrastructural indicators, productivity indicators, and business ease indicators. These indicators have been grouped into categories. For example, GDP, GDP per capita, Exports, ICT Goods Exports, and ICT Goods Imports are grouped into one category of Economic Indicators. The detail description of each indicator along with the unit of measure (UoM) and the corresponding year the data was considered are listed in Appendix A.

It can be observed that closeness to market or suppliers is not considered in this research. The rationale is that these indicators are more specific to the individual companies. As the purpose of this study is to assess the applicability of machine learning algorithms, so more general indicators have been selected. Anyhow, the indicators are a subjective selection and may vary from company to company. The companies can use the indicators that are more relevant to them. Minimum wage, output per worker and health expenditures are collected from International Labour Organization[1]. Global peace index data has been extracted from Vision of Humanity[2]. Whereas, the rest of the data has been collected from the World Bank[3].

The data is collected for a total of 94 countries. Programming language R was used to group all the data sets together. After grouping the data sets, it was found out that there were some countries for which particular indicator values were not available. For the sake of easiness during clustering computations, these values were replaced by the average values of that particular indicator for the rest of the countries. For example, the data for industrial freshwater withdrawal for Latvia was not available, hence the average value of the rest of the countries i.e. 21.2 was assigned to the Latvia.

**Results and findings**
In the first step, the dataset was imported into the Matlab. As all indicators had different units of measure (UoM), hence it was not possible to visualize and compare the data. In the 2nd step, as described in the methodology section, the data was normalized using the mean values. After mean normalization, it was possible to plot and visualize the data. As an example, normalized economic indicators of 94 countries are shown in Figure 2. It can be seen that the developed countries such as China and Germany have relative higher total GDP than the less developed countries like Afghanistan and Nigeria. GDP per Capita is higher for countries like the United States and Ireland. Similarly, the rest of the indicators were also normalized.

**Figure 2.** Normalized economic indicators

In step no. 3 and 4, Principal Component Analysis (PCA) was used to reduce the dimensionality of the data. Initially, data was reduced to two dimensions but the variance retention, as per step number 5 and 6, was lower than 95%. Eventually, Fifteen dimensions gave the variance retention of approximately 95%. Hence, the data was reduced from 24 dimensions to 15 dimensions. In the step number 7, in order to run the k-means algorithm, the number of clusters were to be selected. First, Elbow method was used to plot the sum squared errors against the number of clusters. This method measures the error or variance between cluster centroid and the data points assigned to that cluster. It builds on the idea that the value of the number of clusters should be selected so that adding more number will not give the better modeling of the data (Bholowalia and Kumar, 2014). Figure 3 shows the elbow curve plotted using the reduced dimensionalized data. On y-axis sum of squared errors are shown, whereas on x-axis number of clusters. Figure 3 shows that there is no clear elbow for this data, hence it was impossible to select the exact number of clusters by using this method. Hence, another method called the Calinski-Harabasz method was used to calculate the optimal number of clusters (Calinski and Harabasz, 1974). This method provided the optimal number of clusters as 2. Hence, it was decided to divide the data set into 2 number of clusters.

**Figure 3.** Elbow curve of the dataset

In step no. 8, two random data points were selected from the data set to random initialize the centroids. In steps 9, the distances among centroids and data points were calculated. In step 10, new centroids were assigned by taking the means of the data points that were closest to the previous centroids. In step 11, step 9 and 10 were repeated until satisfactory clusters are achieved. Figure 4 shows the final clusters of dataset after 10 iterations of the k-means algorithm. Red points show the one cluster and the green points show the other.

**Figure 4.** Final clusters after 10 iterations

After obtaining the clustered data, the number of clusters used were validated using Silhouette plots in step no. 12. Silhouette is an efficient method that can be used to decide the number of clusters into which a data set can be divided (Zhou and Gao, 2014). Figure 5 shows 4 different plots of silhouette values. Silhouette function outputs values that are between -1 and 1. According to Silhouette method, the number of clusters that gives more positive and higher values i.e. close to 1 is more efficient than the number of clusters that gives lower and negative values. It can be seen from Figure 5 that 2 number of clusters gives the Silhouette values that are more positive and higher than the rest of the number of clusters. Hence, this also validates the initial assumption of choosing 2 number of clusters in step 7.

**Figure 5.** Silhouette plots for different number of clusters

After the data were clustered, it was easier to visualize and compare the data. Figure 6 shows boxplots of economic indicators of 2 clusters of countries. The center lines inside the box represent medians of the data, whereas edges show the 75th and 25th percentiles of the data. On y-axis, normalized values are shown, whereas on x-axis economic indicators are shown. It can be seen that the cluster 1 countries have the higher GDP, GDP per Capita, Exports, ICT Goods Exports, and ICT Goods Imports as compared to cluster 2 countries. Similarly, boxplots of political, infrastructural, productivity and business ease indicators are also plotted.

**Figure 6.** Medians and percentiles of economic indicators of clustered data
These boxplots were then used to compare the individual values of all the indicators. Table 2 shows the individual scores of both clusters against all indicators. If a higher value is desired for a particular indicator, then "High" is written in the "Desired Value" column and vice versa. The indicator values are compared using boxplots. If desired value of an indicator is "High" and cluster 1 has the higher median value as compared to cluster 2 then "+" is assigned to cluster 1 and "-" is assigned to cluster 2. For example, the desired value of GDP is "High" and as Figure 6 shows the higher median value for cluster 1 than cluster 2, hence "+" is assigned to cluster 1 and "-" is assigned to cluster 2. Similarly, if desired value is "Low" and if cluster 1 has the lower value than cluster 2, then "-" is assigned to cluster 2 and "+" is assigned to cluster 1. For example, Unemployment in Table 2. On the contrary, if median values are very close to each other that the difference is negligible then nothing is assigned to any of the clusters, for example, air freight transport in Table 2.

| Table 2. Scores of two clusters |

After assigning the "+" and "-", individual scores for each category of indicators are calculated. To calculate the individual scores number of "+" are counted for all indicators per one category and then divided by the total number of indicators per one category. For example, there are 4 number of "+" for Economic indicators for cluster 1, whereas the total number of indicators per economic category is 5. Hence, 4 is divided by 5 to get an overall score of 0.80 for cluster 1 countries. Similarly, the rest of the scores are calculated for all indicator categories for both of the clusters. These scores are then plotted in a radar plot and shown in Figure 7. It can be seen that cluster 1 has the higher scores in all of the categories i.e. economic, political, infrastructure, productivity, and business ease than cluster 2. Hence, it can be implied that as a decision maker, it would be reasonable to select cluster 1 countries to be considered for facility location. Now that countries are selected, the decision makers can use any of the other methods to further narrow down the selection and choose a particular country.

**Figure 7. Overall scores of cluster 1 and 2 countries**

**Discussion**

As shown in Table 1, limitations exist for decision frameworks and models in ILD. Some frameworks such as AHP and PERC (Badri, 1999, Viswanadham and Kameshwaran, 2007) are prone to personal subjectivity. Whereas, the other methods such as Stochastic Model (Goh et al., 2007), CFLP Method (Liu et al., 2010), or MIP (Cui et al., 2010) are complex enough that they do not allow a large number of parameters or variables to be considered. Further, most of these methods require prior shortlisting of the alternatives before the implementation.

Moreover, in reviewing the literature, no data was found on the association of machine learning and ILD. This study was set out with the aim of assessing the application of machine learning in facility location. It was hypothesized that k-mean clustering will be useful in applying machine learning in the ILD. The results show that k-mean clustering can be used as a decision framework in facility location decision.

It is interesting to note that the framework developed through this study does not depend on the personal subjectivity. Although the country's indicators or parameters may change based on the preferences of the decision makers, the method does not depend on personal bias. Another important finding of this research is that this method provides a quantitative output that provides a more concrete assessment of the choices under consideration. The most interesting finding of this study is that as k-means clustering can be applied to a very large dataset, hence it can be implied that this framework provides a huge flexibility to include a large number of parameters and variables without any added complexity. Moreover, as discussed in “Literature Review” that almost all of the decision frameworks and models work on predefined alternatives, whereas no method provides a rationale in shortlisting these alternatives. It is interesting to note that this framework efficiently provides initial alternatives to be considered. These alternatives can be used in further methods, such as Gravity Method (Liu et al., 2010), ROI (Biswa et al., 2016), or MIP (Cui et al., 2010) to select a more specific location. Further, as indicators to be considered are flexible, hence this framework may also work effectively at national level location decision.

**Comparison with other frameworks**

As this framework is neither an optimization model nor seeks to be an optimization model this means it is not feasible to compare the results with other ILD frameworks. Moreover, many of the ILD frameworks such as MIP (Cui et al., 2010), ROI, and Dimensional analysis (Biswa et al., 2016) propose an optimization model based on a
single variable such as cost or distance, whereas this framework incorporates multiple variables (24 in this case) that makes it impossible to compare the results on a quantitative basis. In addition, the frameworks that offer multiple criteria to consider in ILD such as AHP (Viswanadham and Kameswaran, 2007), Fuzzy Group Assessment (Lin and Lee, 2007), and Grey Based Multi Criteria Decision Making (Parkash et al., 2015) are prone to personal subjectivity and the results are so dissimilar that these cannot be compared quantitatively. Hence, in this article, 2 fold evaluation of the research is used based on the design evaluations methods outlined by Hevner et al. (2004). First evaluation is the functional testing that has been carried out in this research using the 94 countries dataset and proves the functionality of the algorithm proposed in the Methodology section. Second, descriptive informed arguments are used to prove the superiority of the framework as the nature of the output makes it unfeasible to evaluate quantitative performance. These arguments have been summarized below:

1. This framework provides a quantitative assessment without personal subjectivity
2. The framework can incorporate multiple criteria or variables
3. The ability to incorporate huge datasets makes this framework flexible and easy to use for complex decisions
4. This framework can provide an initial shortlist of countries to consider for later optimization using other frameworks

Implications for practice
The framework is in particular useful for companies that are competing in global markets. Large firms that need to offshore their manufacturing facility or warehouses need to consider a large number of influencing factors, as discussed in the Literature Review section. This framework can be very useful for such companies as it is flexible enough to incorporate a large data set. Moreover, this framework can also be applied by companies at a national level. To use the framework companies will need to go through the following steps:

1. Decide on the factors or indicators that need to be considered for facility location by using expert opinion, delphi approach, or any other suitable method
2. Gather information or data for all of the countries or locations for respective indicators
3. Find the clusters using k-means algorithm as mentioned in the methodology section and Figure 1
4. Score the individual clusters as mentioned in the methodology section and shown in Table 2
5. Find the overall scores of individual clusters as shown in Figure 7 and select the cluster with the highest score

Implications for theory
This study contributes to theory by introducing a novel framework for facility location decision that incorporates multi criteria decision making. This framework helps reducing the personal bias by producing a quantitative output. The flexibility of this framework is the core strength, as the number of criteria or indicators to be considered can be easily increased or decreased without adding the extra complexity to the framework. In addition, this framework can be used to shortlist alternatives to be optimized by other frameworks. In a nutshell, this study provides a fast, flexible, and quantitative framework for practioners to decide on facility location.

Hence it can be interpreted that k-means clustering provides an efficient framework that can be used by decision-makers in facility location. Overall, this is a fast and flexible technique that decision-makers can use as a framework to decide and select a group of alternatives to be considered as potential optimal facility locations.

Conclusion and notes for further research
This article aimed to answer the research question: “How can unsupervised machine learning be used in the international location decision?” To answer the research question, a research was carried out on 24 economic, infrastructural, political, productivity, and business ease indicators of 94 countries. Un-supervised machine learning i.e. k-means clustering was used in the research to assess the outcomes of the analysis. The k-means clustering provided 2 clusters with the optimized correlation among the countries of each cluster. These clusters were then compared and scored individually based on the median values of each indicator. The cluster with the highest score was selected as a group of potential facility locations. Hence, the results of this study suggest that the clustering framework is an efficient technique that can be used in International Location Decision. This framework provides a quantitative output without any personal subjectivity that decision-makers can use to decide on facility locations. This framework is flexible enough that it can adjust and incorporate several parameters according to the preferences of the decision makers. It can be very useful in selecting or shortlisting initial potential locations in the ILD. Industry professionals can use this framework in strategically selecting the facility locations. Further, it also encourages
Limitations
There are several limitations that have been encountered during the analysis. For example, during the data collection, many indicator values were not available for some of the countries. Hence, average values were used for the countries for which no data was available. This can alter the final results, but the methodology or approach still remains valid. Further, only general 24 indicators are used in the analysis, whereas addition or negation of some indicators can alter the final results. Moreover, the types of indicators are subject to personal preferences and decision makers can use the indicators that are more relevant to them. Some important indicators, such as closeness to market and suppliers that result in lower transportation costs are not considered in this analysis due to unavailability of the data. Further, these parameters are company specific and can easily be added to the analysis. Keeping all the limitations stated above, the approach is very much dependent on the data. Hence, one may get a different number of clusters and results than derived in this analysis.

References
Ferdows, K. (1997), Making the Most of Foreign Factories.
Jayaram, A. (2016), "Lean six sigma approach for global supply chain management using industry 4.0 and IIoT", in Aradhya, V. N. M. (Ed.), Institute of Electrical and Electronics Engineers Inc.


Schorr, H. and Rappaport, A. (1989), "Innovative applications of artificial intelligence".


Wang, L., He, J. and Xu, S. (2017), "The application of industry 4.0 in customized furniture manufacturing industry", in Matec web of conferences 100.


Zhang, L. Y., Sun, Y. and Ma, J. (2010), "A decision support system of logistics location", IEEE.


Appendix A