Pamukkale Univ Muh Bilim Derg, 28(4), 625-631, 2022

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Assessing the heterogeneity of social connectedness index via quantile regression mixture model

Kantil regresyon karma model ile sosyal bağlılık endeksinin heterojenliğinin incelenmesi

Tolga KURTULUŞ[1], *Serpil KILIÇ DEPREN*[2*]

[1]Graduate School of Science and Engineering, Yildiz Technical University, Istanbul, Turkey.
tolgakurtulus95@gmail.com

[2]Department of Statistics, Faculty of Arts & Science, Yildiz Technical University, Istanbul, Turkey.
serkilic@yildiz.edu.tr

Received/Geliş Tarihi: 08.06.2021 Revision/Düzeltme Tarihi: 06.10.2021 doi: 10.5505/pajes.2021.16446
Accepted/Kabul Tarihi: 25.10.2021 Research Article/Araştırma Makalesi

Abstract

This study aims to visualize a network of Social Connectedness Index (SCI) in Organization for Economic Co-operation and Development (OECD) countries and then explores the importance of socio-demographic, economic, religion, and distance metrics between countries on SCI using a non-parametric test. The final dataset is aggregated from 3 different data sources: Worldbank, OECD, and Facebook. Drawing upon a data set from Facebook Inc. is used to visualize and understand the network structure among OECD countries. Furthermore, the aggregated dataset used in this study is the first usage of Quantile Regression Mixture Models (QRMIX) to determine factors affecting SCI. As a result of the QRMIX model, 4 clusters are identified in different quantiles where the impact of independent factors are differentiated. Based on the variable importance analysis, almost the least important variable at the lower level of SCI value is religion while it is the second most important factor at the highest level of SCI value. SCI mostly shows up as strong relationships between countries with residents of similar ages and education levels where using common language and having same religions, as well. Also, based on the literature review, it is shown that countries with a higher proportion of similar connections to other countries have more positive economic connections among OECD countries. Thus, given the variable importance of SCI for different subgroups of based on SCI quantiles, this study suggests that different action plans about improving import-export and other financial transactions for the country pairs might be created. To sum up, according to different social connection power of OECD countries, this study can help policymakers.

Keywords: Quantile regression mixture models, Social connectedness index, Non-Parametric method, Social network analysis.

Öz

Çalışmanın amacı, Ekonomik Kalkınma ve İşbirliği Örgütü (OECD)'ne üye ülkelerin arasındaki sosyal bağlılık yapısının görselleştirilmesi ve ülkelerdeki sosyo-demografik, ekonomik, din ve resmi dil faktörlerinin Sosyal Bağlılık Endeksine (SCI) etkisinin parametrik olmayan bir yöntemle incelenmesidir. Çalışmadaki veriseti 3 farklı veri kaynağından alınmıştır: Dünya Bankası, OECD ve Facebook. Facebook firmasından alınan veriler OECD üye ülkeleri arasındaki sosyal ağ yapısının görselleştirilmesi ve özelliklerinin anlaşılması için kullanılmıştır. Ek olarak veri kaynaklarından alınan tüm verilerin birleştirilmesiyle elde edilen kümüle veri seti, SCI'ya etki eden faktörlerin Kantil Regresyon Karma Modeller (QRMIX) ile belirlenmesi amacıyla kullanılmıştır. QRMIX sonucunda, SCI'ya etkisi farklılaşan 4 küme belirlenmiştir. Önem derecesi analizine göre, SCI'nın düşük seviyelerinde ülkenin resmi dini önemi en düşük faktör iken SCI'nın yüksek seviyelerinde en önemli ikinci faktördür. Benzer yaş, eğitim seviyesi, dil ve dine mensup ülkeler arasında güçlü bir sosyal ağ yapısı olduğu gösterilmiştir. Ayrıca, literatürdeki çalışmalarda benzer sosyal ağ gücüne sahip ülkelerarasında ekonomik olarak da güçlü bir bağ olduğu gösterilmiştir. Bu çalışmada da, SCI değerinin farklı kantillerine göre bağımsız faktörlerin öneminin değiştiği, bu sebeple farklı sosyal bağlantılara sahip ülkelerin ithalat-ihracat ve finansal işlemler gibi göstergeleri için farklı aksiyon planları oluşturulabileceği önerilmiştir. Sonuç olarak, bu çalışma farklı bağlantı gücüne sahip OECD ülkeleri için politika yapıcılarına yardımcı olabilir.

Anahtar kelimeler: Kantil regresyon karma modeller, Sosyal bağlılık endeksi, Parametrik olmayan yöntemler, Sosyal ağ analizi.

1 Introduction

The physical distance between people is not as important as in the last few decades any longer because social networks let us reaching new ideas, information, people and, institutions in a quicker, easier, and cheaper way. That is why social networks have a great impact on both our social and business life in terms of improving social skills and employment opportunities. Furthermore, besides the impact of social networks on individuals' life, it shapes the global society in terms of social mobility and travel, migration, and political behaviors, as well [1]-[4]. Moreover, a strong social connection between countries is one of the crucial factors affecting countries' financial metrics such as foreign direct investments, the volume of imports and exports [5]. Since the social network structure is highly correlated with different financial and non-financial metrics, it is important to understand which factors have a significant impact on this network structure across countries.

In the extant literature, it is hard to find a robust metric that measures the social connectedness between countries. As Facebook.com platform is the world's largest online social networking service which has almost 3 billion accounts, the "Social Connectedness Index" (SCI) metric proposed by Facebook is used to understand the behavioral and social closeness of nations with their Facebook connections. Also,

*Corresponding author/Yazışılan Yazar

users who shared their locations on their Facebook profile page are in this metric where their device and connection information are also included to infer the exact location of a user.

Formally, the $Social\ Connectedness_{i,j}$ ($SC_{i,j}$) [6] between two locations i and j is defined as:

$$SC_{\mathrm{i,j}} = \frac{\mathrm{FB\_Connections_{ij}}}{\mathrm{FB\_Users_i} * \mathrm{FB\_Users_j}} \quad (1)$$

In Equation (1), FB_Users_i and FB_Users_j are the numbers of Facebook users in locations i and j, and $FB\_Connections_{ij}$ represents the total number of Facebook friendship connections between individuals in the two locations. Dividing by the numbers of Facebook users makes us understand the more potential friendship links between countries with more Facebook users. As a result, it can be said that if $SC_{i,j}$ is twice as large, a Facebook user in the country i is about twice as likely to be connected with a given Facebook user in region j. Within the usage of SCI, it can be assumed that Facebook friendship links may be related to international economic indicators and sociological similarities of countries. At this point, the aim is to find out if there is a relation either significant or insignificant to make further analyses.

There are limited studies about determining factors affecting the social connectedness between countries or defining the relationship between SCI and countries' metrics in the extant literature. In the study of Bailey et al. [5], the correlation between the number of passenger train trips, which is named travel, and social connectedness is analyzed using regression analysis. Distance between regions, rail time, and drive time are used as independent variables that are thought to have a significant impact on SCI. As a result, it is found that the higher SCI between regions causes a higher number of passenger train trips between them. Furthermore, in the study, it is analyzed whether the similarity between the two country pairs in terms of social and financial metrics have a statistically significant impact on SCI. As a result of the study, it is revealed that the higher similarity between countries the higher SCI is seen as well.

In the recent study of Kuchler, Russel, and Stroebel in 2020 [7], it is focused on the relationship between Corona Virus (Covid-19) and SCI using time series analysis. The number of Covid-19 cases per 10k is used as the dependent variable. As a result of the study, it is revealed that there is a strong positive correlation between SCI and the number of Covid-19 cases. In the study of Kuchler et al. in 2020 [8], the impact of financial factors on SCI is analyzed in detail. In the study following research questions are examined "how institutional investors' portfolio decisions are influenced by SCI". In this context, different factors, which are firms' institutional ownership, firm valuation, and firm liquidity, are taken into consideration to measure the impact of these factors on SCI. Also, there are some studies aimed to determine the nexus between social network and gross domestic product, foreign direct investment, and other financial indicators in the literature. Rauch (2001) is studied about the relationship between social networks and international trade [9]. Researchers are focused on two research problems which are "Do networks have a significant impact on efficiency of trade?" and "Can networks improve importance for international trade over time?". Similar to the study of Rauch, the relationship between trade and social networks is examined in the study of Wagner, Head, and Ries (2003) [10]. Furthermore, researchers are worked on whether business activity is affected by the structure of social networks [11], [12]. Furthermore, there are some studies analyzing the relationship between foreign direct investment and social networks in the literature [13]-[15].

The rest of this study is organized as follows. In Section 2, mathematical preliminaries of SNA and QRMIX are given. Then, Section 3 describes the dataset and empirical analysis. The visualization of SCI of countries and analysis results are evaluated in Section 4. Finally, Section 5 includes the discussion and conclusion part of the study.

1.1 Research focus

The study focuses on visualizing the network structure of OECD countries using social connections among them and understand the impact of the background characteristics affecting different SCI levels via the Quantile Regression Mixture Model. Since the QRMIX model produces detailed information on different levels of SCI, the results let policy-makers create more effective action plans on how to increase SCI between countries.

Thus, they can improve their international flows because SCI is highly correlated to international trade, people flow from country to country, import, and export [6].

To determine the factors affecting SCI among countries, we addressed the following research questions:

1. What are the characteristics of the social network in OECD countries? Can this network be visualized via Social Network Analysis?
2. What are the factors that have a statistically significant impact on SCI?
3. Are the impacts of the factors affecting SCI differentiated at the different levels (quantiles) of SCI?

2 Material and method

Before analyzing how networks can be constituted and the factors affecting SCI in countries that differ according to quantiles, we present the general information about the methods.

2.1 Social network analysis

Social Network Analysis (SNA) has an applicable technique to map, measure, and analyze patterns and types of relationships between actors such as individuals, financial organizations, or locations. SNA has contributed to quantitative metrics of several qualitative insights [16]-[18]. To better understand a social network, we define a graph G in a graph theory which is denoted by $G = \{V, E\}$ where V is a set of nodes (vertex) and E is a set of edges (links) herein. Each node represents actors that have a visual representation in a network map by nodes. The nodes are individuals and the edges are any social connection between these nodes. There are different types of the graph: directed and undirected. While the undirected graph is used to present only symmetrical relations, the directed graph is also used to show both symmetric and asymmetric relations.

The most commonly used network metrics can be employed to study the patterns and structures of a social network. These can be summarized in Table 1 [18]-[22]:

Table 1. Summary of network metrics.

Metric	Description
Centrality	It is determined the positions and importance of actors regarding how close they are to the center of the action within a network. When all nodes in the network have similar centrality scores, this network is said to be decentralized. The higher centrality scores the node is, the less similarity it is to demonstrate in network structure,
Degree Centrality	It is defined as the number of edges connecting a specific node to the other nodes in the network. Since only the connections of a node with its neighbors are considered in the degree centrality, it is a local metric. Nodes with a high degree of centrality are taken into account to have an impact on many nodes and can connect with their neighbors,
Betweenness Centrality	It is not concerned with how far a node is from other nodes, but whether that node is on the shortest path between two other nodes. Betweenness centrality is considered as a global measure that is the most complex metric among the centralities. According to this centrality, the most central node is the node with the highest value.

2.2 Quantile regression mixture models

Finite mixture models (FMM) have been widely used for modeling unobserved (latent) sub-populations heterogeneity by consisting of data visualization, clustering, and classification. Quantile regression allows conditional distribution of the response variable, such as the median, on the covariates [15]. A semi-parametric mixture of quantile regressions model explains the relationships between the response variable and the covariates without any parametric assumption on the error densities. If the error probability density function (pdf) is asymmetric, it can be preferred median or other quantiles as a measure of central location. When the population has heterogeneous sub-populations for modeling nonlinear regression relationships, the traditional regression models are not sensitive to outliers and heavy-tailed distributions. To overcome this situation, Quantile Regression Mixture Models (QRMIX) is used in this study. It is well known that the proposed QRMIX improves to be the robustness of the model in the presence of weak conditions. As compared to a traditional mixture of regression models, QRMIX yields more reliable parameter estimates to outliers and extreme values by fitting varying conditional quantile functions.

QRMIX has been applied in many different fields such as economic and financial researches [23], [24], clinical trials [25], psychological research [26], and educational research [27]. However, all of these researches provide model-based clustering on the conditional distribution of the response variable for each error component.

The univariate mixture models may be written as

$$Z \sim \sum_{j=1}^{m} \pi_j g(z - \mu_j) \tag{1}$$

where the error density function f is symmetric about zero. Suppose that (X, Y) has a multivariate random vector with the distribution. We denote by $h(x)$ the marginal density of $X \in \mathbb{R}^p$. Then, the semi-parametric mixture of the regression model is

$$f_x(y) = f(y|x, \quad \theta, G) = \sum_{j=1}^{m} \pi_j \, g(y - x^T \beta_j) \tag{2}$$

where $f(.)$ is the conditional density of $Y|X = x$ and g is completely unspecified with a median equal to zero. π_j and β_j are the parameters of interest. Let $\theta = (\pi_1, \pi_2, \dots, \pi_m, \beta_1, \beta_2, \dots, \beta_m, g)$ denote the vector of parameters.

$$p_{ij} = P(Z_{ij} = 1|x_i, \theta) = \frac{\pi_j g(y_i - x_i^T \beta_j)}{\sum_{l=1}^{m} \pi_l g(y_i - x_i^T \beta_l)} \tag{3}$$
$$\mathrm{i} = 1, \dots, n; j = 1, \dots, m$$

Z_{ij} is an unobserved indicator that the i^{th} observation belongs to the j^{th} mixture component. Then, one finds through posterior probabilities of the Bayesian approach. These responsibilities or membership probabilities play an important role in mixture models. Firstly, the probabilities of class membership are updated given the current parameter estimates for initializing the algorithm, and secondly, after convergence of that algorithm. A weight matrix $P = (p_{ij})$ can be seen as a final decision to which class each observation comes from. They are used for clustering and classification goals.

To estimate the unknown parameters in this model, we find through maximum likelihood estimation defined as follows:

$$L(\theta) = \prod_{i=1}^{n} \left(\sum_{j=1}^{m} \pi_j \, g(y_i - x_i^T \beta_j) \right) \tag{4}$$

For later use, the popular Expectation-Maximization (EM) algorithm is generally used to estimate θ by iterating, starting from some value θ, between the E-step and M-step [28]-[30].

We omit a detailed technical description of how the algorithm is conducted within the background as this has been covered in the extant literature [31].

3 Dataset and empirical analysis

The 5-step procedure is used in this study. Data gathering and data manipulations are the first and the second steps of the study. The third step is the data visualization via the social network analysis. Performing the QRMIX model is the fourth step and interpreting the results is the fifth step of the study.

In the first step, raw-data is gathered from Facebook Inc., OECD, and World Bank aggregated [32]-[34]. In the second step, the country-to-country dataset of SCI is visualized using social network analysis, and network characteristics are given in detail. Since the dependent variable, which is SCI, is design as "from Country-A to Country-B", independent factors are calculated as "factor of Country-A minus the factor of Country-B" to measure the distance (or in other words measure of similarity) between countries in the third step. QRMIX approach is performed with the assumption of having a different impact of independent factors at different levels of SCI,

in the fourth step. Finally, in the fifth step, results are interpreted using model outputs and variable importance analysis results.

Only OECD countries are included in the analysis in the study. Since the Israel dataset is not available, the dataset used in the study consists of 36 OECD countries. There are 1.260 observations in the dataset because of the "from country to country" data design. The Independent factors used in the study are as follows [5], [35]:

- The factor "Distance" is defined as the distance between two countries in miles,
- The factor "Education" is the percentage of the population at least completed lower secondary education,
- The factor "Age" is the median age of the country.
- The factor "GDP per Capita" is the GDP per capita with the current international dollar,
- The factor "Unemployment" is defined as the ratio of the unemployed to the total labor force,
- The factor "Religion" is the official religion of the country. It is flagged as 0 if the official religion of both countries is the same, 1 otherwise,
- The factor "Language" is the official language of the country. It is flagged as 0 if the official language of both countries is the same, 1 otherwise.

The dependent variable of the model is SCI, but natural logarithm transformation is used in the analysis. RStudio tool is used to perform the Social Network Analysis and QRMIX model. The "network", "igraph", "tidyverse", and "tidygraph" packages are used for SNA, and the "qrmix" package is used for QRMIX in the study.

4 Results and discussion

SNA is a useful tool to understand and visualize the connection among countries. Overall network structure among countries is given in Figure 1, where the thicker the line between the two countries, the stronger the social connection between these countries. Furthermore, descriptive statistics of each country in the network, which are degree, closeness, and betweenness centralities, are examined to understand the basics of the network structure.

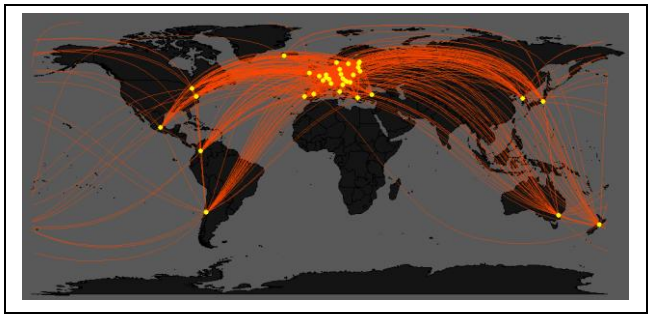

Figure 1. The social network structure of OECD countries based on SCI.

For being able to receive valuable findings from this study, a world map plot is prepared using SCI values and networks between OECD countries. Each color in Figure 1 represents the Social Connectedness Index value within the relation of countries with the pre-selected country, which is Turkey. Colors on the countries are distributed on a continuous scale and getting darker as long as the SCI index value increases, decreases otherwise.

In Table 2, it is given that Switzerland, Ireland, Iceland, Luxemburg, and United Kingdom have the highest degree of centrality statistics in the network. Switzerland has 56 connections in total if the network structure is assumed as a non-directed structure. The closeness centrality statistics are almost the same for each country in the network. According to the betweenness centrality measure, Sweden is the most central node in the network, followed by Canada, New Zealand, Germany, and Austria. Furthermore, the minimum, average, and maximum values of the network are given in Table 2. Based on the network averages, one country has 29 connections out of 36 countries in the network on average.

Table 2. Top 5 Countries in the network by the different centrality measures.

#	Degree	Closeness	Betweenness
1	Switzerland (56)	United Kingdom (0.0137)	Sweden (180)
2	Ireland (54)	Italy (0.0137)	Canada (122)
3	Iceland (52)	Ireland (0.0136)	New Zealand (96)
4	Luxembourg (52)	Spain (0.0136)	Germany (84)
5	United Kingdom (50)	Australia (0.0136)	Austria (82)
Min.	2	0.0008	0
Avg.	29	0.0128	31
Max.	56	0.0137	180

Centrality statistics are given in the parenthesis.

SNA mapping characteristics for the country "Turkey" on OECD member countries on Europe continent is given in Figure 2(a) where North-South America continents are given in Figure 2(b) and on Asia-Australia continents is given in Figure 2(c).

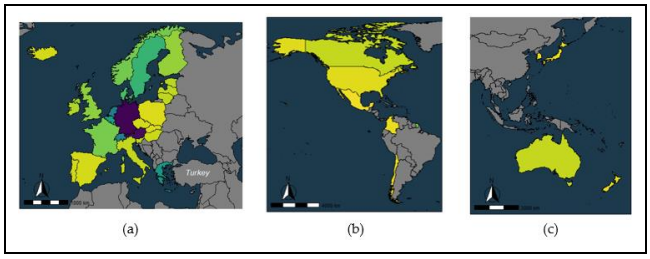

Figure 2. Social Connectedness Index of Turkey among OECD countries located in Europe (a): North-South America, (b): and Asia-Australia (c).

As a gathering from Figure 2(a), it is clear to notice that there is high social connectedness relation from Turkey to Germany and Austria where many ethnic Turkish families living in these countries. Besides, Belgium, Netherlands, and Switzerland also have high social connectedness relation with Turkey, which could be predicted as either for labor or ethnic reasons. Lastly, as Turkey and Greece have historical date back connections and relations from the past, their social connection is also high.

As it is shown in Figure 2(b), it is easy to notice colors are not as dark as in Figure 2(a) colors. This is because Turkey does not have high social connectedness relations with Canada, the USA, Mexico, Colombia, and Chile. However, we can gather that

Canada has the highest social connectedness index among North-South American countries and Guyana has a darker color, which is considered as a part of France.

The first noticeable gathering from Figure 2(c) is that there are few countries are located in the plot. These countries are South Korea, Japan, Australia, and New Zealand. It is claimed that colors are not as dark as in Figure 2(a), which may be the reason for the low levels of social connectedness relations of Turkey with these countries.

After visualization of the dataset, the QRMIX approach is performed for 3, 4, and 5 clusters. In this study, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and minimum sample size of the clusters are taken into consideration for the selection of the best model.

According to Table 3, 3-cluster, 4-cluster, and 5-cluster models are compared regarding model selection criteria. Although the 5-cluster model has the lowest RMSE and MAE statistics, its minimum sample size is lower than 100, which can be led to misleading results in the QRMIX model. Thus, the results of the 4-cluster model are interpreted in the study.

Table 3. RMSE and MAE statistics of the model with 3, 4, and 5 clusters.

# of Cluster	RMSE	MAE	Min. Sample Size
3	0.181	0.146	214
4	0.142	0.115	116
5	0.112	0.087	98

The quantiles, adjusted R^2 values, number of observations, coefficients, and standard errors of the 4-cluster model are given in Table 4.

Table 4. 4-Cluster model output.

Model	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Quantile (τ)	*14%*	*50%*	*82%*	*96%*
R^2	*90.3%*	*93.5%*	*88.0%*	*87.1%*
# of obs.	*388*	*498*	*258*	*116*
Variables	β coef. (S.E.)	β coef. (S.E.)	β coef. (S.E.)	β coef. (S.E.)
Intercept	4.59* (0.03)	4923.000* (34.70)	5228.00* (37.20)	5867.00* (102.50)
Distance *(x1000)*	-0.124* (0.004)	-85.410* (2.058)	-53.440* (2.186)	-73.810* (5.828)
Education	-0.008* (0.001)	-8.102* (0.523)	-8.910* (0.834)	-4.041 (2.120)
Age	-0.024* (0.002)	-46.400* (1.604)	-49.680* (2.770)	-37.730* (7.529)
GDPPP *(x1000)*	0.006* (0.000)	7.304* (0.342)	6.840* (0.535)	9.173* (1.409)
Unemployment	-0.009* (0.003)	-5.090* (1.806)	-6.000* (2.817)	-22.940* (7.305)
Religion	-0.103* (0.017)	-98.260* (11.940)	-112.300* (17.960)	-384.700* (55.770)
Language	-0.661* (0.031)	-637.500* (33.920)	-554.900* (33.590)	-514.700* (101.100)

The QRMIX approach is provided 4 clusters that are statistically significantly differentiated from each other in terms of beta coefficients. The τ values are 14%, 50%, 82%, and 96% from Cluster 1 to 4, respectively. Also, R^2 values are quite high for each cluster. Moreover, the number of observations is enough to interpret the model. On contrary to the studies in the literature, performing the QRMIX model shows that the impact of factors affecting SCI is differentiated according to the level of SCI. Thus, classical regression or time-series analysis could be useful but it is not enough for determining the robust analytical model.

In all clusters, all β coefficients of independent factors are negative, except GDP per capita, which means that the increasing dissimilarity between countries causes to decrease in social connectedness. In general, these findings are similar to the study of Bailey et al. in 2020 [5], especially in the distance, education, age, religion, and language factors. On the other hand, it is revealed that the negative impacts of unemployment and religion on social connectedness are increasing from Custer-1 to Cluster-4, which is an important finding that is not given in the studies in the literature. Also, the impact of distance between countries and language is relatively higher in Cluster-2 (median level of SCI). On contrary to the study of Bailey et al. [5], the higher similarity of unemployment between countries, the higher SCI occurs as well.

Although the coefficients of factors in Cluster-1 are relatively lower than other Clusters, it supports the assumption of the study in terms of the sign of the coefficients. As a result, it is proved that if countries have similar education levels, median age, GDP per capita, unemployment rate, religion, and language then the social connectedness index between countries would be relatively high. Furthermore, the impacts of the factors affecting the social connectedness index are significantly higher at the higher level of the social connectedness index, where are Cluster-2, 3, and 4.

The plot of the actual over the predicted values obtained from the QRMIX model, which is given in Figure 3, is used as a goodness of fit criteria as well. In this plot, it is expected that each dot is on (or around) the 45-degree-line for the model with high prediction accuracy. Based on Figure 3, almost all dots are on (or around) the 45-degree-line. Thus, it can be said that the model has a high level of accuracy. Moreover, in each cluster, the R^2 value is higher than almost 90%, which is relatively higher than the studies in the literature. This finding also supports the QRMIX approach produces more robust results rather than classical approaches.

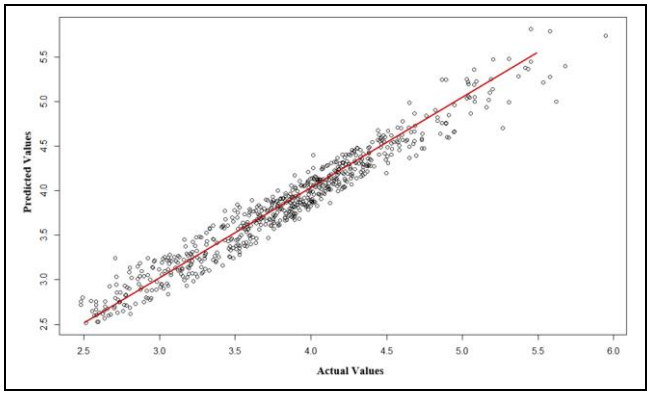

Figure 3. The actual vs predicted values.

The variable importance level of factors affecting the social connectedness index in each cluster is given in Figure 4.

According to Figure 4, the distance between countries is the most influencing factor affecting SCI in Cluster-1 while the importance level of other factors is too close to each other. Moreover, the variables' importance levels are differentiated from cluster to cluster. The language and education factors have lost their importance from Cluster-1 to Cluster-5 while the importance of distance between countries, GDP per capita, and unemployment is almost stable in each cluster. In other words, language and education are getting lost their importance on SCI at the high level of SCI, especially above the median level (Cluster-2). The importance level of age is more important in Cluster 2 and 3. On the other hand, religion is the second most important factor in Cluster-4, which is the highest level of SCI, while it is ranked the last in terms of importance in other clusters. These results show that the QRMIX approach is suitable to model SCI since the importance of factors is differentiated in each cluster.

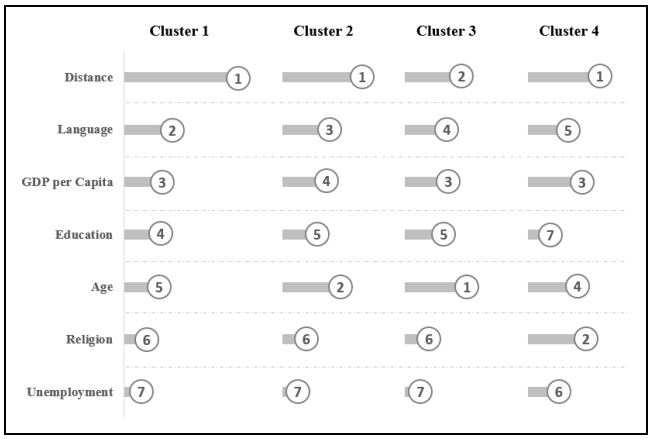

Figure 4. Variable importance for each cluster.

5 Conclusions

The aggregated dataset collected from different data sources is used to understand the social connection among OECD countries in this study. After visualization of the network structure on a map, the basic characteristics of the network are calculated. Finally, the QRMIX approach is performed to determine the factors affecting the different levels (quantiles) of SCI. On contrary to the literature review, the distance between countries, education level, median age, gross domestic product per capita, unemployment ratio, religion, and language factors are used as independent factors while SCI is used as the dependent variable in the QRMIX model. As a result of the analysis, it is revealed that the impact of independent factors on SCI is differentiated based on the level of SCI. This finding is the main contribution to the literature. Thus, policy-makers should be created different action plans about improving import-export and other financial transactions for the country pairs having different social connection power. Using this information, potential countries can be determined in terms of demographics characteristics. Furthermore, similar to studies in the literature, it is shown that similar age, education level, unemployment ratio, religion, and language increase the social connection between countries in this study.

This study has some limitations as well as being a reference study for future research. First of all, the analysis performed with only OECD countries can be extended to a more comprehensive study in which all countries are included. Also, the analysis part can be improved by adding different variables that can affect SCI.

6 Author contribution statements

In the scope of this study, Tolga KURTULUŞ and Serpil KILIÇ DEPREN were contributed equally in terms of the formation of the idea, the design, the literature review, assessment of obtained results, supplying the dataset used, and examining the results were contributed.

7 Ethics committee approval and conflict of interest statement

There is no need to obtain permission from the ethics committee for the article prepared.

There is no conflict of interest with any person / institution in the article prepared.

8 References

[1] Abraham A, Hassanien AE, Snasel V. *Computational Social Network Analysis: Trends, Tools and Research Advances*. Dordrecht, Netherlands, Springer, 2010.

[2] Bailey M, Cao R, Kuchler T, Stroebel J, Wong A. "Measuring Social Connectedness". National Bureau of Economic Research, Cambridge, USA, 23608, 2017.

[3] Bailey M, Kuchler T, Russel D, State B, Stroebel J. "The Determinants and Effects of Social Connectedness in Europe". Center for Economic Studies and Ifo Institute (CESifo), Munich, Germany, 8310, 2020.

[4] Baker M. "The impact of social networking sites on politics". *The Review: A Journal of Undergraduate Student Research*, 10(1), 72-74, 2009.

[5] De Brun A, McAuliffe E. "Social network analysis as a methodological approach to explore health systems: a case study exploring support among senior managers/executives in a hospital network". *International Journal of Environmental Research and Public Health*, 15(3), 1-11, 2018.

[6] Drakulich KM. "Social capital, information, and perceived safety from crime: the differential effects of reassuring social connections and vicarious victimization". *Social Science Quarterly*, 96(1), 176-190, 2015.

[7] Erisen E, Erisen C. "The effect of social networks on the quality of political thinking". *Political Psychology*, 33(6), 839-865, 2012.

[8] Facebook Inc. "Social Connectedness Index". https://dataforgood.fb.com/tools/social-connectedness-index/ (01.02.2020).

[9] Rauch JE. "Business and social networks in international trade". *Journal of Economic Literature*, 39(4), 1177-1203, 2001.

[10] Wagner D, Head K, Ries J. "Immigration and the trade of provinces". *Scottish Journal of Political Economy*, 49(5), 507-525, 2002.

[11] Danis WM, Clercq DD, Petricevic O. "Are social networks more important for new business activity in emerging than developed economies? An empirical extension". *International Business Review*, 20(4), 394-408, 2011.

[12] Combes P-P, Lafourcade M, Mayer T. "The Trade-Creating effects of business and social networks: evidence from France". *Journal of International Economics*, 66(1), 1-29, 2005.

[13] Paniagua J, Korzynski P, Mas-Tur A. "Crossing borders with social media: Online social networks and FDI". *European Management Journal*, 35(3), 314-326, 2017.
[14] Gao T. "Ethnic Chinese networks and international investment: evidence from inward FDI in China". *Journal of Asian Economics*, 14(4), 611-629, 2003.
[15] Forstenlechner I, Mellahi K. "Gaining legitimacy through hiring local workforce at a premium: The case of MNEs in the United Arab Emirates". *Journal of World Business*, 46(4), 455-461, 2011.
[16] Hua J, Huang M, Huang C. "Centrality metrics' performance comparisons on stock market datasets". *Symmetry*, 11(7), 1-15, 2019.
[17] Hunter DR, Young DS. "Semiparametric mixtures of regressions". *Journal of Nonparametric Statistics*, 24(1), 19-38, 2012.
[18] Javed M, Tuckova Z, Jibril AB. "The role of social media on tourists' behavior: an empirical analysis of millennials from the czech republic". *Sustainability*, 12(18), 1-19, 2020.
[19] Kalantan ZI, Einbeck J. "Quantile-Based estimation of the finite cauchy mixture model". *Symmetry*, 11(9), 1-19, 2019.
[20] Kılıç Depren S, Gökalp Yavuz F. "The network analysis of the domestic and international air transportation structure of Turkey". *Mugla Journal of Science and Technology*, 4(2), 148-155, 2018.
[21] Kılıç Depren S. "Determination of the factors affecting students' science achievement level in Turkey and Singapore: An application of quantile regression mixture model". *Journal of Baltic Science Education*, 19(2), 247-260, 2020.
[22] Kuchler T, Russel D, Stroebel J. "The Geographic Spread of Covid-19 Correlates with the Structure of Social Networks as Measured by Facebook". National Bureau of Economic Research, Cambridge, USA, 26990, 2020.
[23] Kuchler Li Y, Peng L, Stroebel J, Zhou D. "Social Proximity to Capital: Implications for Investors and Firms". National Bureau of Economic Research, Cambridge, USA, 27229, 2020.
[24] Koenker R, Bassett G. "Regression quantiles". *Econometrica*, 46, 33-50, 1978.
[25] McLachlan GJ, Peel D. *Finite Mixture Models*. 1st ed. New York, USA, John Wiley & Sons, 2000.
[26] Organization for Economic Co-operation and Development. "Data". https://data.oecd.org/ (01.02.2020).
[27] Raza N, Shahzad SJH, Tiwari AK, Shahbaz M. "Asymmetric impact of gold, oil prices and their volatilities on stock prices of emerging markets". *Resources Policy*, 49, 290-301, 2016.
[28] Richard JW, Ching-Ray Y, Emir B, Zou KH, Cabrera J. "A Comparison and Integration of Quantile Regression and Finite Mixture Modeling". *Joint Statistical Meeting*, Boston, USA, 2-7 August 2014.
[29] Small ML. "Racial differences in networks: do neighborhood conditions matter?". *Social Science Quarterly*, 88(2), 320-343, 2007.
[30] Smithson M, Shou Y. "CDF-Quantile distributions for modelling random variables on the unit interval". *British Journal of Mathematical and Statistical Psychology*, 70, 412-438, 2017.
[31] World Bank. "DataBank". http://databank.worldbank.org/data/home.aspx (01.02.2020)
[32] Weare C, Loges WE, Oztas N. "Email effects on the structure of local associations: a social network analysis". *Social Science Quarterly*, 88(1), 222-243, 2007.
[33] Wu Q, Yao W. "Mixtures of quantile regressions". *Computational Statistics & Data Analysis*, 93, 162-176, 2016.
[34] Yum S. "Social Network Analysis for Coronavirus (COVID-19) in the United States". *Social Science Quarterly*, 101(4), 1642-1647, 2020.
[35] Zivkovic R, Gajic J, Brdar I. "The impact of social media on tourism". *Sinteza*, 1, 758-761, 2014.