Advances in Industrial Engineering, Winter 2022, 56(1): 1-14
DOI: 10.22059/AIE.2022.333903.1814

RESEARCH PAPER

A New Hybrid Data Mining Technique to Forecast the Greenhouse Gases Emissions

Hosein Khodami, Reza Kamranrad [*], Ehsan Mardan

Department of Industrial Engineering, Semnan University, Semnan, Iran.

Received: 13 November 2021, Revised: 16 February 2022, Accepted: 16 March 2022

Abstract

The expansion of industrial activities and the unnecessary growth of cities have increased the concentration of greenhouse gases, including carbon dioxide in the atmosphere. Mostly, CO2 emissions are caused by the consumption of different forms of energy and the combustion of all types of fuels, especially fossil fuels. The development of data mining techniques that lead to accurate prediction of CO2 emissions is very useful in deciding the preventive measures and appropriate policies in this area. Most studies in this field are limited to models that do not compare different techniques and features and only examine the effect of economic factors and fossil fuel consumption on CO2 emissions. The aim of this study is to identify a combination of significant features as well as to select the best technique to predict CO2 emissions. For this purpose, a huge dataset containing various features was obtained from the IEA database. A new hybrid method for predicting CO2 emissions was developed, and then results were compared with proposed data mining techniques including ANN, KNN, GLE, Linear-AS, and Regression. Also, a combination of significant features and the best techniques for predicting CO2 emissions were identified. The results of the proposed method show that by clustering the database and then implementing prediction techniques, the error could be substantially reduced. Also, in order to predict future observations, first, they are placed in appropriate clusters with the Discriminant Analysis technique and then they are predicted with the appropriate forecasting technique. It was found that the proposed hybrid technique, which is a combination of K-Means, Linear-AS and Discriminant Analysis, is most accurate in this case.

Keywords:
Data Mining;
Energy Consumption;
Greenhouse Gases Emission;
Statistical Analysis;
Global Warming

Introduction

Climate is changing as a result of human activities, and evidence suggests that this trend continue in the future. Most scholars who have studied the consequences of climate change believe that the consequences of this phenomenon will be detrimental to the interests of human society. The phenomenon of global warming has raised debate among opponents and supporters of this issue as to whether greenhouse gases contribute to the phenomenon. So, proponents of the issue take the consequences of this phenomenon very seriously; and in contrast, opponents believe that there isn't sufficient evidence of the effect of greenhouse gas emissions on global warming. Some opponents also argue that controlling the emission of this gas will prevent economic and industrial growth, however, many scientific evidences and statistical studies indicate that increasing greenhouse gas emissions, including CO_2, are one of the key causes of global warming [1]. Alam et al. argued that excessive energy consumption, especially fossil

[*] Corresponding author: (R. Kamranrad)
Email: r.kamranrad@semnan.ac.ir

fuels, contributes to environmental pollution in order to achieve economic growth goals; so that, CO_2 emissions, one of the major contributors to air pollution, are the result of the use of fossil fuels in the manufacturing, commercial, service and household sectors [2].

In recent years, due to the increase of population and industrialization of societies, the CO_2 emission trend is increasing; therefore, concerns about the adverse consequences of this have also increased. Therefore, many studies have investigated the factors influencing CO_2 emission and many strategies have been introduced to reduce its emission. In the meantime, predicting the amount of CO_2 emissions is important in providing the right direction for policies adopted by governments and relevant organizations. The CO_2 emission trend, from 1965 to 2017, is shown in Fig. 1 and, as can be seen, its increasing emission trend can be a source of concern. Fig. 2 also shows the average share of each country in emissions over the past 52 years. As is clear, the main CO_2 emitters are mostly industrialized and developed countries, indicating the relevance of industrialization to CO_2 emissions; because these countries have to increase energy consumption to maintain continued economic and industrial growth. Energy production from fossil fuels emits far more CO_2. But this does not mean that other types of energy production such as nuclear, hydro and renewable energies play no role in CO_2 emissions, rather indirectly reduce or sometimes increase CO_2 emissions. Figs. 3 to 5 illustrate the energy consumption trends for fossil fuels including petroleum, natural gas, coal, as well as renewable energies, hydropower and nuclear energy. As it is clear, due to factors such as population growth and industrialization of societies, energy consumption is increasing. Data mining uses various techniques and data analysis tools to discover patterns and relationships hidden in huge databases [3]. Specifically, data mining can play an important role in controlling greenhouse gas emission and understanding its impact on the environment [4]. Predicting the amount of greenhouse gas emission, including CO_2, using data mining techniques has helped greatly in finding the right policies to prevent adverse consequences.

Most previous studies have examined the association of CO_2 emissions with various factors, including economic issues, population growth, industrialization of communities, consumption of fossil fuels in different sectors, and so on. Also, some studies have predicted CO_2 emission rates for the coming years, using various time series techniques. However, it should be noted that different countries have differing CO_2 emissions patterns over many years due to technical issues such as energy efficiency. Therefore, this study, using a set of data mining techniques, predicts the amount of emission of this gas for each country and each year according to the features presented in the next sections. This study attempts to identify the best combination of available features by combining data mining techniques, as well as selecting the best method from the proposed techniques to predict CO_2 emissions.

The research dataset was collected from the International Energy Agency (IEA) website. The IEA is an independent intergovernmental organization that publishes energy data annually. The reason for choosing this dataset is that it is more complete, accurate, and more common than other datasets. In this study, five prediction techniques including ANN, KNN, GLE, Linear-AS, and Regression as well as a new hybrid proposed method combining Clustering, Discriminant Analysis, and Linear-AS techniques in the prediction model used for IEA data. Also, in order to compare the present techniques, the mean square error index, which is common in many studies, was used to select the best technique. The remainder of this study is organized as follows: Section 2 provides an overview of the subject literature and related studies. Section 3 presents the data and features obtained from the dataset. In Section 4, the research methodology is presented which is constructed using data mining techniques to discover useful patterns among the data. Section 5 also deals with the conclusions of the study.

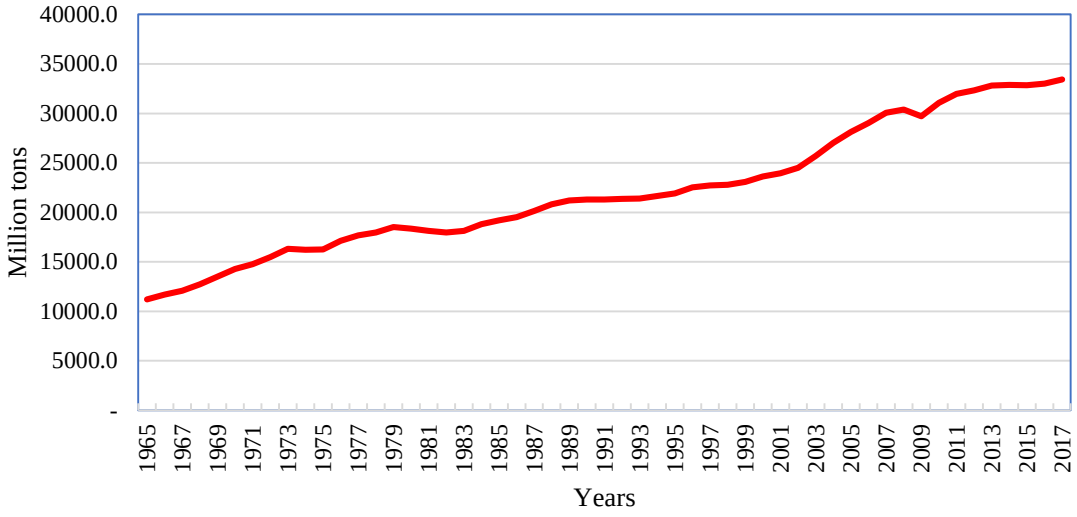

Fig. 1. CO_2 emission trends

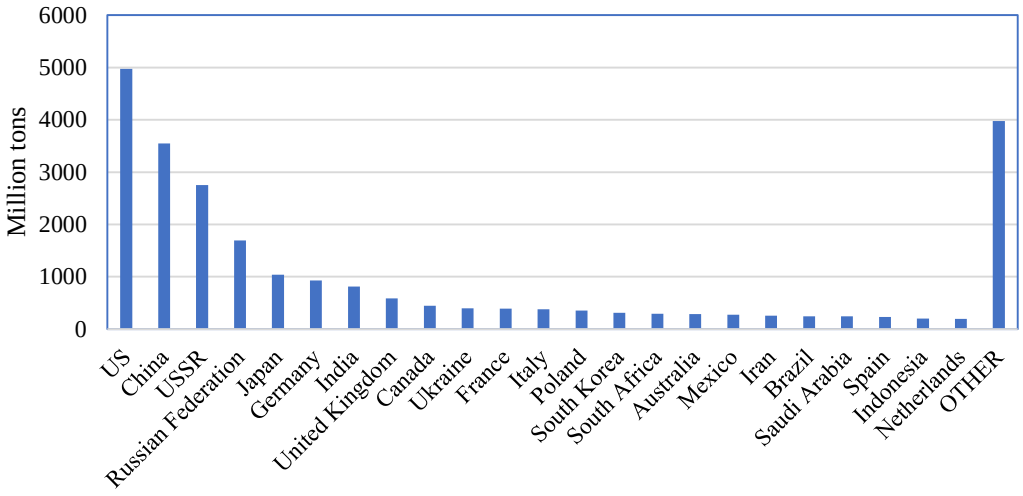

Fig. 2. Average CO_2 emissions by different countries in the last 52 years

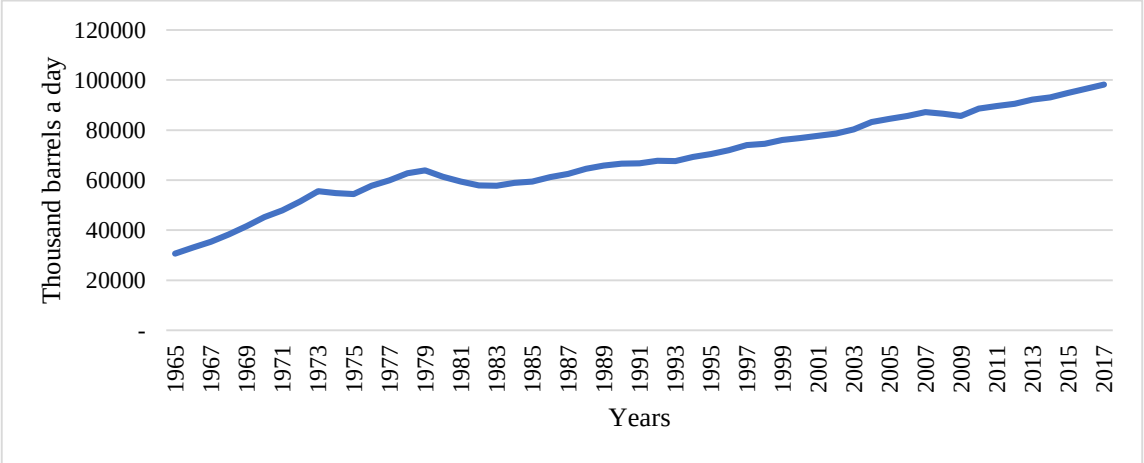

Fig. 3. Oil consumption trends in the world

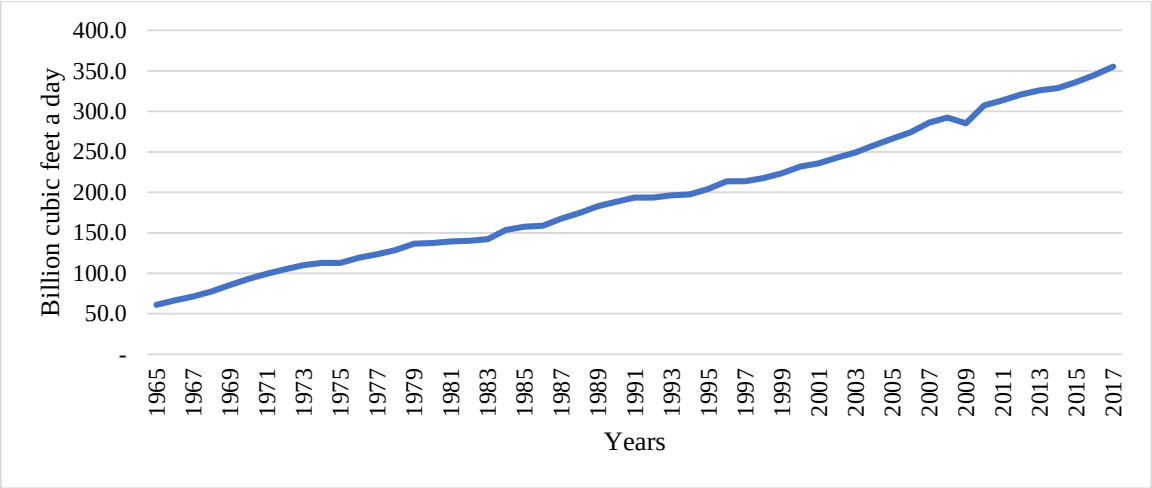

Fig. 4. Gas consumption trends in the world

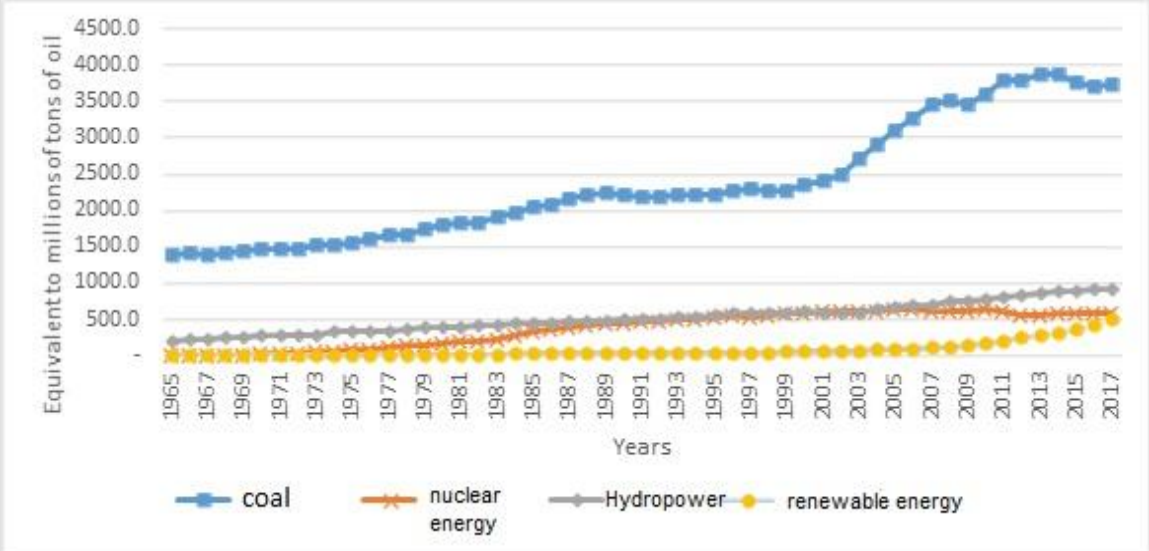

Fig. 5. Trend of energy consumption of coal, nuclear, hydro, renewable in the world

Literature review

This section reviews studies on the factors influencing greenhouse gas emissions including CO_2 and their proposed techniques for forecasting. As expected, due to the importance of identifying the factors that contribute to CO_2 emission as well as anticipation of strategies to counter them, extensive studies have been carried out, which will be discussed below. Holtz-Eakin and Selden [1], investigated the relationship between economic growth in communities and CO_2 emissions, using the data extracted from the panel data to prove that there is the law of diminishing returns between CO_2 emissions and GDP per capita. Knapp and Mookerjee [5] investigated the relationship between population growth and CO_2 emissions using the Granger Causality Test and cointegration model during the years 1880 to 1989. Begum et al. [6], examined the relationship between CO_2 emissions and energy consumption and economic growth and population in Malaysia, so that in their study, the experimental results of the ARDL approximation showed that CO_2 emissions increased exponentially with further GDP growth during the period 1970 to 1980 to 2009, as a result of the Environmental Kuznets Curve (EKC) is not approved for Malaysia. Furuoka studied the relationship between CO_2 emissions and development [7]. For this purpose, he used three different methods including cross-sectional regression, stacked regression and threshold regression to investigate this relationship. Jalil and

Mahmud studied the relationship between CO_2 emissions and per capita income for China using the ARDL method from 1975 to 2005 [8]. He and Richard examined the relationship between CO_2 emissions and economic growth in Canada between 1948 and 2004 [9]. Tol et al. investigated the long-term relationship between energy consumption and CO_2 emissions in the United States during the 1980s and 1980s [10]. The main results of their study show that the intensity of CO_2 emissions increases with the increase of fossil fuels and population growth, economic growth and electricity consumption growth are also factors influencing CO_2 emissions. Halicioglu examined the relationship between CO_2 emissions, income, and foreign trade for Turkey, whose results suggested that there was a long-term relationship between CO_2 emissions and energy and income [11]. Lotfalipour et al. predicted the CO_2 emission for Iran using the Grey System and Autoregressive Integrated Moving Average and also compared the two methods using the RMSE, MAE and MAPE criteria [12]. Hamzacebi and Karakurt presented a gray prediction model to predict CO_2 emissions in Turkey and predicted the amount of greenhouse gas emissions for Turkey by 2025 [13]. Nyoni and Bonga developed and tested an ARIMA (2,2,0) forecasting model using the Box-Jenkins ARIMA approach, using annual CO_2 emissions data for India [14].

Marjanović et al. investigated the relationship between CO_2 emissions and economic growth and predicted the GDP based on CO_2 emissions using the Extreme Learning Machine (ELM) technique [15]. After modeling, they also compared the results with the results of genetic techniques and artificial neural networks. Liu et al. Analyzed the factors driving production-based CO_2 emissions in Beijing. Their results show that production-based CO_2 emissions in Beijing have jumped from 12.78 million tons in 1980 to 45.91 million tons in 2015, but the growth rate of CO_2 emissions has slowed from 2010 to 2015 [16]. Zheng et al. studied the major contributors to energy-related CO_2 emissions in China [17]. Kunda and Phiri studied the trend of CO_2 emissions from fossil fuels for major industries using data from Zambia and they used the SMOreg algorithm for time series to predict the amount of emission of this gas [18]. Kumar and Muhuri with a new learning-based approach called transfer, predicted different countries' GDPs using CO_2 emissions and compared the results of the proposed model with 4 methods of Generalized Regression, Neural Network, Extreme Learning Machine, and Support Vector Regression [19]. Yu and Du using data from the Chinese provincial panel, examined the impact of technology innovation on CO_2 emissions and used the logistic equation to predict CO_2 emissions during the 2016-2030 period [20]. Hosseini et al. predicted CO_2 emissions for Iran considering two possible scenarios by 2030, using multiple linear regression and multiple variable regression techniques [21]. Li et al. predicted the amount of CO_2 emissions based on the Cointegration Theory for China's data, taking into account factors including GDP per capita, urbanization level, energy intensity, and total energy consumption [22]. Wen and Cao predicted CO_2 emission trends in China, identified 18 primary indices using gray relation analysis, and then modeled it with a proposed algorithm called ICSO-SVM [23]. Table 1 summarizes the useful information available in the literature review studies.

As is well known in the literature, existing studies focus more on examining the impact of factors such as economic conditions, fossil fuels, and various industries on CO_2 emissions and presenting a model to predict it. However, so far no study has been conducted on the prediction of CO_2 emissions due to different energy consumption levels at a macro level for countries around the world for many years. In this research, while developing a new hybrid technique, CO_2 emissions for each country have been predicted for many years using the International Energy Agency (IEA) dataset. The results are also compared with the results of several common data mining techniques.

Table 1. A summary of the literature review

Existing research	Purpose of the study	Research Methods
Holtz-Eakin and Selden [1]	Investigate the relationship between economic growth in communities and CO2 emissions	multiple regression
Knapp and Mookerjee [5]	Investigate the relationship between population growth and CO2 emissions	Granger Causality Test and cointegration model
Jalil and Mahmud [8]	The relationship between CO2 emissions and per capita income	ARDL Approximation
Begum et al. [6]	Examined the relationship between CO2 emissions and energy consumption and economic growth and population	ARDL Approximation
Furuoka [7]	Study of the relationship between CO2 emissions and development	Three different methods including cross-sectional regression, stacked regression, and threshold regression
Lotfalipour et al. [12]	Predicted the CO2 emission for Iran	Grey System and Autoregressive Integrated Moving Average
Nyoni and Bonga [14]	Predicted the CO2 emission for India	Box-Jenkins ARIMA
Marjanović et al. [15]	Investigate the relationship between CO_2 emissions and economic growth and predicted the GDP	Extreme Learning Machine (ELM) technique
Kunda and Phiri [18]	Study of the trend of CO2 emissions from fossil fuels for major industries	SMOreg algorithm
Kumar and Muhuri [19]	predict different countries' GDPs using CO2 emissions	Generalized Regression, Neural Network, Extreme Learning Machine and Support Vector Regression
Wen and Cao [23]	predict CO_2 emission trends in China	ICSO-SVM - gray relation analysis -

The data set

The most comprehensive and commonly used reference in the field of energy is the International Energy Agency's (IEA) vast database, with most energy balance sheets and related studies selecting their data from this database. The IEA is an independent intergovernmental organization established within the framework of the Organization for Economic Co-operation and Development (OECD) in the wake of the 2007 oil crisis and has its secretariat in Paris. The organization was initially founded in response to physical disruptions in oil supply, also as a source of information on the international oil market and other forms of energy. The data obtained from this dataset covers 92 countries from 1965 to 2017. Evidence suggests that CO_2 emissions have a direct relationship with the amount of energy used, such as the amount of oil, natural gas, coal, nuclear energy, hydropower, and renewable energy. Each of the factors mentioned is either directly or indirectly involved in CO_2 emissions; For example, the increasing energy consumption of hydropower instead of coal is expected to reduce CO_2 emission. Different countries also have different conversion rates in CO_2 emissions for energy production due to different combustion facilities, technologies, and standards. For example, developed countries in the field of fossil fuels have Euro standards to reduce air pollution, but developing countries do not have these standards or have similar and weak standards. A similar argument holds for the feature of the year that these standards did not matter in previous years,

but in recent years, attention has been paid to these standards and other constructive measures to reduce air pollution, which causes the combustion compounds, one of which is CO_2, change. Therefore, year and country were also selected as features from this dataset. This database contains 4876 objects (92 countries and 53 years) that are presented in Table 2, each with a brief description. The first attribute is related to the name of the country and the second attribute is related to its year and the rest attributes relate to the amount of consumption of different energy sources as described in Table 2. The CO_2 emission variable is also a predictor variable.

Table 2. A summary of the features of the IEA dataset

features	Description	Variable type
Country	The country name, which in this study includes 92 countries	**Nominal**
Year	Year: From 1965 to 2017	**Nominal**
Oil consumption	Oil consumption in thousand of barrels per day	**Numerical**
Natural gas consumption	Natural gas consumed in billion cubic feet per day	**Numerical**
The amount of coal consumed	Coal consumption equal to millions of tons of oil per day	**Numerical**
Hydropower consumption	Hydropower consumption equal to millions of tons of oil per day	**Numerical**
nuclear energy consumption	nuclear energy consumption equal to millions of tons of oil per day	**Numerical**
Consumption of renewable energy	renewable energy consumption equal to millions of tons of oil per day	**Numerical**
CO_2 emissions	CO_2 emissions per million tons	**Numerical**

Methodology

In this study, a new hybrid method is used to predict CO_2 emission according to the relevant variables. The RMSE is also used to evaluate the model and the results are compared with several other data mining techniques. Prior to modeling, the Kruskal-Wallis test was used to examine whether country and year variables had a statistically significant effect on CO_2 emissions. The new hybrid method described in Section 4-3 is a combination of K-Means, Linear-AS, and Discriminant Analysis techniques. Also, available techniques for comparison with the proposed method are ANN, LINEAR, KNN, REGRESSION, and LINEAR-AS techniques. The best model is selected by comparing the RMSE index value for each technique. In addition, the best combination of variables available for prediction is also identified simultaneously.

The research method of this research is a combination of data mining techniques to predict the amount of greenhouse gas emissions. The process of the combined method is such that first the existing data are classified into different groups by clustering method and Co_2 emissions are then predicted for each group using data mining forecasting techniques including ANN, LINEAR, KNN, REGRESSION, and LINEAR-AS. The prediction error for each prediction method is then compared for a different number of clusters. Finally, a forecasting method will be selected that has less error for the existing clusters. So far, the number of optimal clusters and the best forecasting technique have been identified; Now, to predict future data, first using the DA-technique, the appropriate cluster for the existing data is identified, and then again, the optimal forecasting technique is implemented in the corresponding clusters and prediction is performed. Continuing the research in Section 4-1, the problem data are preprocessed for data mining operations and then in Section 4-2, the statistical significance of year and country variables is analyzed using the Kruskal-Wallis test. Then in Section 4-3 the proposed hybrid technique is described in detail and modeling is performed in Section 4-4. Fig. 6 also illustrates the research process.

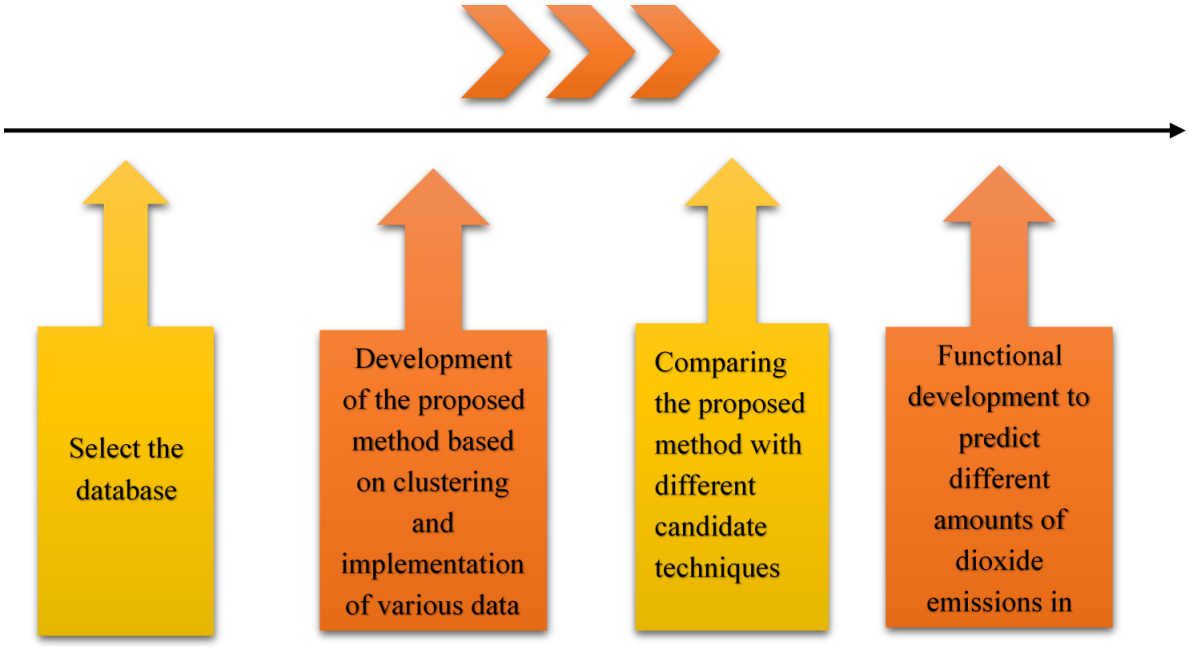

Fig. 6. Research process diagram

Preparation and pre-processing of data

This dataset contains missing data for some countries, especially in previous years, including 601 observations that were deducted from the dataset and left a total of 4275 observations. Then, for the remaining observations, for each country name variable, a code was assigned to each country, coding for 92 existing countries from 1 to 92. For example, Canada was given the number 2. In this data set, after deleting the Outlier, there are 4275 records of which 70% (2992) of them have been used for training and 30% (1283) of them have been used for testing.

Kruskal-Wallis test for year and country variables

The Kruskal-Wallis test for significant differences in the factors influencing CO_2 emissions was used to determine whether country and year influenced the factors introduced for CO_2 emissions. For this purpose, the Kruskal-Wallis test was used to determine the influence of country and year type on observations on the presented model by which the variables are at a 95% confidence level. 92 countries in the study were assigned 92 distinct codes from 1 to 92, and 53 different years were numbered with codes 1 to 53, and the test results are described in Table 3. According to the results, the chi-square test statistic for both variables is very high and the P-VALUE variables are 0. This implies that it can be surely concluded that the type of country and year of observation has a significant statistical effect on CO_2 emissions. Therefore, due to the influence of these two variables on CO_2 emission, these two variables will not be eliminated in the study process to predict the amount of CO_2.

Table 3. Kruskal-Wallis test on variables

Variables	chi-square	p-value
Year	247.12	0
Country	3315.37	0

Proposed hybrid technique

In many cases, the observations in the datasets may not be sufficiently coordinated, which can lead to a large error in the prediction process. There are various reasons for data inconsistency with regard to data type and field of study. For example, the data studied in the field of energy and CO_2 emissions, for reasons such as differences in energy consumption patterns and CO_2 emissions for different countries as well as different years, reduce the coordination of observations. Therefore, in some cases, the use of a clustering pattern causes observations that are more similar and coordinated to one cluster and reduce the error rate by applying the prediction model. Also, all modeling has been implemented in IBM SPSS Modeler 18 software, which is the reason that this software is selected, its comprehensiveness and ease of use.

The proposed method combines three data mining techniques, including K-Means, Linear-AS, and Discriminant Analysis. According to Wagstaff et al. [24], the K-Means algorithm is a method used to automatically cluster data into K categories. The function of this technique is to first select the centers of the first clusters and then proceed as follows:

1. Each dj observation is assigned to its closest cluster
2. Each Cj cluster center is updated in order to average the cluster observations.

On the other hand, the Linear-AS technique is used to predict each cluster. Linear models are relatively simple to understand and can generally be built into a dataset faster than other models such as artificial neural networks or decision trees [25]. LINEAR-AS is applicable when the IBM SPSS Modeler 18 is connected to the relevant analytics services, and this modeling is usually more accurate than Linear.

After clustering, by applying the LINEAR-AS technique to the data, and using trial and error, we try to determine the optimal number of clusters so that the average error of the clusters is minimized. In fact, the value of k, which represents the value of the clusters, must be chosen such that from n, the following relation state exists:

$$Min\ (mean\{MSE_1, MSE_2, \dots, MSE_k\}) \qquad k = 1, \dots, n$$

It should be noted that by implementing clustering, usually from a threshold value thereafter, increasing the value of the clusters increases the error because by decreasing the observations within the clusters the accuracy of the prediction model decreases so that the values of K Usually a small number. After selecting the K value and running the model, in order to predict the new observations, it is first necessary to determine which clusters belong to each observation. This decision should be based on the maximum likelihood of the new observation being similar to the observations of existing clusters. For this purpose, the Discriminant Analysis (DA) technique is used. Discriminant Analysis is one of the statistical techniques used in machine learning and pattern recognition to find the linear combination of properties that best separates two or more classes of objects [26].

The DA finds coordinate axes (K-1 standard coordinates, K denotes the number of classes) that best separate the categories. These linear functions are uncorrelated, separating k-1 spaces through an n cloud of data that best separates K groups. Now to retrieve clusters from this technique, each class is assumed to be a single cluster, so new observations can be referenced to the cluster to which they belong by examining the equations obtained by the DA technique. For this purpose, Minitab 17 statistical software is used, so that for each observation a cluster number is considered as its class number. After the implementation of the DA model, a membership equation is obtained for the number of available clusters (K numbers). By substituting the values of the various variables for a new observation, a cluster is selected whose equation gives more numerical value.

In order to evaluate the proposed technique and compare it with other prediction techniques, it is tested for the existing database and the results are presented in the following sections.

Selecting the best technique and optimal combination of variables

Modeling with Data Mining techniques and Feature Selection

In this section of the study, the results of the proposed hybrid technique are compared with the results of the five types of data mining techniques that are most used in quantitative prediction. These techniques include REGRESSION, KNN, GLE, ANN, and LINEAR-AS. As the modeling is performed for each technique, the best combination of variables is also identified. In this study, for each of the techniques, different combinations of variables (at least 4) are tested. The best combination of variables is a combination that minimizes model prediction error. In general, the total number of combinations achievable for a set of 8 is n state, where, in this study, a single subset of variables cannot be less than 4. Therefore, all the subsets obtained by calculating the total number of combinations are as follows:

$$2^n - \left(\frac{n!}{1!(n-1)!}\right) - \left(\frac{n!}{2!(n-2)!}\right) - \left(\frac{n!}{3!(n-3)!}\right) - 1$$

Where n is the total number of variables equal to 8. So by placing the value in the above expression, it becomes clear that 163 different states must be examined for each technique. Table 4 shows the best combination of variables for all five data mining techniques, with respect to the error index. The results show that the best modeling method with the least error is the LINEAR-AS method, with all eight variables including country, year, oil, gas, coal, hydroelectric, nuclear, and renewable. The RMSE index value for this technique is 12.71 which is significantly lower than the other techniques.

Table 4. RMSE index for the present techniques

Technique	RMSE	Attribute / Variable Combination
ANN	7.79	Country, year, oil consumption, gas consumption, coal consumption, Hydropower consumption, nuclear energy consumption,
KNN	6.23	Country, year, oil consumption, gas consumption, coal consumption, Hydropower consumption, nuclear energy consumption, renewable energy
GLE	5.56	Country, oil consumption, gas consumption, coal consumption, Hydropower consumption, nuclear energy consumption
REGRESSION	8.26	Country, year, oil consumption, gas consumption, coal consumption, Hydropower consumption, nuclear energy consumption, renewable energy
LINEAR-AS	3.56	Country, year, oil consumption, gas consumption, coal consumption, Hydropower consumption, nuclear energy consumption, renewable energy

Development of the proposed hybrid technique

In the previous section, the error index obtained for the LINEAR-AS technique, given that the figures are large, shows an acceptable value, but since the error index is generally an undesirable index, it is used to reduce it. Using the KMEANS technique, the existing data are clustered and LINEAR-AS is used for each cluster of prediction. Given that the observations of each cluster have the highest magnitude of harmony and coordination, the prediction error is expected to decrease. For this purpose, non-hierarchical clustering is performed for 2 to 5 clusters. The reason that the upper limit of clustering 5 is considered is that as the number of clusters (k) increases, clusters with very low numbers of observations are created. This causes

the prediction error for these clusters to be very high and thus to increase the mean error; therefore, the model need not be tested at k = 5. Table 5, which shows the modeling results, indicates that the best number of clusters (k*) for the available data is 2. As can be seen from Table 5, the RMSE decreased by 32.02% on average using clustering, which is a very significant value.

Table 5. Summary of Proposed Model Results

Technique	The number of observations per cluster	The best combination of features in the cluster	Number of clusters = 2	RMSE	Average of RMSE
LINEAR-AS	2073	Country, Year, Oil consumption, Gas consumption, Coal consumption, Hydropower consumption, Nuclear energy consumption, Renewable energy	Cluster 1	2.10	2.42
	2202		Cluster 2	2.73	

Technique	The number of observations per cluster	The best combination of features in the cluster	Number of clusters=3	RMSE	Average of RMSE
LINEAR-AS	2031	Country, Year, Oil consumption, Gas consumption, Coal consumption, Hydropower consumption, Nuclear energy consumption	Cluster 1	2.70	3.21
	2133		Cluster 2	2.14	
	111		Cluster ۳	4.78	

Technique	The number of observations per cluster	The best combination of features in the cluster	Number of clusters=4	RMSE	Average of RMSE
LINEAR-AS	1993	Country, Year, Oil consumption, Gas consumption, Coal consumption, Nuclear energy consumption, Renewable energy	Cluster 1	3.63	5.94
	1871		Cluster 2	3.52	
	337		Cluster 3	6.07	
	74		Cluster 4	10.57	

Technique	The number of observations per cluster	The best combination of features in the cluster	Number of clusters=5	RMSE	Average of RMSE
LINEAR-AS	548	Year, Oil consumption, Gas consumption, Coal consumption, Hydropower consumption, Nuclear energy consumption	Cluster 1	6.93	7.96
	1742		Cluster 2	4.31	
	1286		Cluster 3	4.60	
	663		Cluster 4	5.80	
	36		Cluster 5	18.17	

As is clear, the best way to predict CO_2 emissions is to combine clustering (Where k = 0) and linear-as. Also, the best combination of variables is introduced as the 8 variables. Now for the new observations, one must first identify the cluster to which the observation belongs, and then perform the linear-as prediction. DA technique is used for this purpose. So that the number of clusters in the clustering was used as the group number in DA. In this way, it is possible to identify suitable clusters for new observations and predict them for each cluster accordingly. The results show that this technique has very high accuracy and low error for this series of data and also the model output is shown in Table 6, which indicates 99.5% accuracy for predicting the type of cluster, for new observations.

Table 6. DA technique results

	Cluster 1	Cluster 2
Number of observations per cluster	2073	2202
Proportion correct	99.4%	99.6%

Therefore, first, in order to identify the clusters corresponding to each observation, we use Eqs. 1 and 2 obtained by the differentiation analysis technique via Minitab17. Then proceed to predict CO_2 emissions for observations, in accordance with the function trained in the LINEAR-AS technique for each cluster.

$$Y_1 = -35285 + 36 \times YEAR + 2 \times OIL - 3 \times GAS - 3 \times HYDROELECTRICITY - 45 \times RENEWABLES \quad (1)$$

$$Y_2 = -36176 + 36 \times YEAR + 2 \times OIL - 3 \times GAS - 3 \times HYDROELECTRICITY - 45 \times RENEWABLES \quad (2)$$

Conclusions and suggestions

Conclusions

This study presents a novel method that combines K-Means, Linear-AS, and Discriminant Analysis techniques to predict CO_2 emissions. Also, the proposed method was compared with 5 common prediction data mining techniques including REGRESSION, KNN, GLE, ANN, and LINEAR-AS using the RMSE error index. The results show that the proposed combination technique can reduce 32.02% error in this particular case. In fact, by clustering the data, the similarity and uniformity of the data in each cluster increases, and as a result, the LINEAR-AS prediction model makes a better estimation. It should also be noted that the choice of the number of clusters depends on the type and number of data. So from a k value onwards, the mean error of the clusters increases with decreasing data within the clusters. Therefore, the optimal value of k is determined very quickly by trial and error. In addition, the best combination of features available from the possible scenarios was identified. It is suggested to use this hybrid technique in datasets that, for the same reasons, are not uniform for different reasons, and have little intrinsic similarity. The variables in this study were energy-related features that, by testing all possible combinations of them, it was found that the least error was obtained for all eight model variables. Given that the energy consumption pattern of each country depends on its policies, it is recommended that before applying the policies, using this model, which is highly accurate, to check the CO_2 emissions.

Future research suggestions

The proposed hybrid method can be used in different datasets that have data inconsistencies. It is also possible to compare the results with other techniques not considered in this study to evaluate the effectiveness of this method. Also, by combining other methods, new methods can be developed that suit different datasets. It is worth mentioning that it is possible to predict the amount of CO_2 emissions for the coming years by using time series techniques to predict the amount of different energy consumed.

As a suggestion for future studies, we can mention the use of the group learning method. In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Unlike a statistical ensemble in statistical mechanics, which is usually infinite, a machine learning ensemble consists of only a concrete finite set of alternative models but typically allows for a much more flexible structure to exist among those alternatives.

Forecast of CO_2 emissions for Iran in the next 5 years

In this section, according to the data from the energy balance sheet obtained from the official website of the Iranian Statistics and Information Network (www.isn.moe.gov.ir), we have attempted to predict the amount of CO_2 emissions in Iran. For this purpose, we first assign the obtained data to the appropriate clusters with DA, and in this particular case, which includes the data for 2018 to 2023, all records were assigned to cluster 1. Then, using the linear AS technique, we predict the CO_2 emissions, the results of which are given in Table 7.

Table 7. CO_2 emission forecast results for Iran until 2023

year	Estimated value for CO_2 emissions
2018	680
2019	697
2020	813
2021	805
2022	731
2023	905

References

[1] ALAM, S., FATIMA, A. & BUTT, M. S. 2007. Sustainable development in Pakistan in the context of energy consumption demand and environmental degradation. *Journal of Asian Economics,* 18**,** 825-837.

[2] BEGUM, R. A., SOHAG, K., ABDULLAH, S. M. S. & JAAFAR, M. 2015. CO2 emissions, energy consumption, economic and population growth in Malaysia. *Renewable and Sustainable Energy Reviews,* 41**,** 594-601.

[3] FOX, J. 1997. *Applied regression analysis, linear models, and related methods*, Sage Publications, Inc.

[4] FURUOKA, F. 2015. The CO2 emissions–development nexus revisited. *Renewable and Sustainable Energy Reviews,* 51**,** 1256-1275.

[5] HALICIOGLU, F. 2009. An econometric study of CO2 emissions, energy consumption, income and foreign trade in Turkey. *Energy Policy,* 37**,** 1156-1164.

[6] HAMZACEBI, C. & KARAKURT, I. 2015. Forecasting the energy-related CO2 emissions of Turkey using a grey prediction model. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects,* 37**,** 1023-1031.

[7] HE, J. & RICHARD, P. 2010. Environmental Kuznets curve for CO2 in Canada. *Ecological Economics,* 69**,** 1083-1093.
[8] HOLTZ-EAKIN, D. & SELDEN, T. M. 1995. Stoking the fires? CO2 emissions and economic growth. *Journal of public economics,* 57**,** 85-101.
[9] HOSSEINI, S. M., SAIFODDIN, A., SHIRMOHAMMADI, R. & ASLANI, A. 2019. Forecasting of CO2 emissions in Iran based on time series and regression analysis. *Energy Reports,* 5**,** 619-631.
[10] JALIL, A. & MAHMUD, S. F. 2009. Environment Kuznets curve for CO2 emissions: a cointegration analysis for China. *Energy policy,* 37**,** 5167-5172.
[11] KARGUPTA, H., GAMA, J. & FAN, W. The next generation of transportation systems, greenhouse emissions, and data mining. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010. 1209-1212.
[12] KLECKA, W. R., IVERSEN, G. R. & KLECKA, W. R. 1980. *Discriminant analysis,* Sage.
[13] KNAPP, T. & MOOKERJEE, R. 1996. Population growth and global CO2 emissions: A secular perspective. *Energy Policy,* 24**,** 31-37.
[14] KUMAR, S. & MUHURI, P. K. 2019. A novel GDP prediction technique based on transfer learning using CO2 emission dataset. *Applied Energy,* 253**,** 113476.
[15] KUNDA, D. & PHIRI, H. 2017. An Approach for Predicting CO2 Emissions using Data Mining Techniques. *International Journal of Computer Applications,* 172**,** 7-10.
[16] LI, X., SONG, Y., YAO, Z. & XIAO, R. 2018. Forecasting China's CO2 Emissions for Energy Consumption Based on Cointegration Approach. *Discrete Dynamics in Nature and Society,* 2018.
[17] LIU, Z., WANG, F., TANG, Z. & TANG, J. 2020. Predictions and driving factors of production-based CO2 emissions in Beijing, China. *Sustainable Cities and Society,* 53**,** 101909.
[18] LOTFALIPOUR, M. R., FALAHI, M. A. & BASTAM, M. 2013. Prediction of CO2 emissions in Iran using Grey and ARIMA models. *International Journal of Energy Economics and Policy,* 3**,** 229-237.
[19] MAIMON, O. & ROKACH, L. 2005. Data mining and knowledge discovery handbook.
[20] MARJANOVIĆ, V., MILOVANČEVIĆ, M. & MLADENOVIĆ, I. 2016. Prediction of GDP growth rate based on carbon dioxide (CO2) emissions. *Journal of CO2 Utilization,* 16**,** 212-217.
[21] NYONI, T. & BONGA, W. G. 2019. Prediction of CO2 Emissions in India Using ARIMA Models. *DRJ-Journal of Economics & Finance,* 4**,** 01-10.
[22] TOL, R. S., PACALA, S. W. & SOCOLOW, R. 2006. Understanding long-term energy use and carbon dioxide emissions in the USA.
[23] WAGSTAFF, K., CARDIE, C., ROGERS, S. & SCHRÖDL, S. Constrained k-means clustering with background knowledge. Icml, 2001. 577-584.
[24] WEN, L. & CAO, Y. 2019. Influencing factors analysis and forecasting of residential energy-related CO2 emissions utilizing optimized support vector machine. *Journal of Cleaner Production***,** 119492.
[25] YU, Y. & DU, Y. 2019. Impact of technological innovation on CO2 emissions and emissions trend prediction on 'New Normal'economy in China. *Atmospheric Pollution Research,* 10**,** 152-161.
[26] ZHENG, X., LU, Y., YUAN, J., BANINLA, Y., ZHANG, S., STENSETH, N. C., HESSEN, D. O., TIAN, H., OBERSTEINER, M. & CHEN, D. 2020. Drivers of change in China's energy-related CO2 emissions. *Proceedings of the National Academy of Sciences,* 117**,** 29-36.