Energy Conversion and Management 65 (2013) 185–197

Contents lists available at SciVerse ScienceDirect

Energy Conversion and Management

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

# A data mining approach: Analyzing wind speed and insolation period data in Turkey for installations of wind and solar power plants

Ilhami Colak [a,*], Seref Sagiroglu [b], Mehmet Demirtas [a], Mehmet Yesilbudak [c]

[a] Department of Electrical and Electronics Engineering, Faculty of Technology, Gazi University, Ankara, Turkey
[b] Department of Computer Engineering, Faculty of Engineering, Gazi University, Ankara, Turkey
[c] Department of Electronics and Automation, Vocational College of Haci Bektas Veli, Nevsehir University, Nevsehir, Turkey

ARTICLE INFO

*Article history:*
Received 26 March 2012
Received in revised form 14 June 2012
Accepted 7 July 2012
Available online 16 October 2012

*Keywords:*
Data mining
Wind speed
Insolation period
Similarity
Feasibility and numerical analyses
Wind and solar power plant installations

ABSTRACT

Wind and solar power plant installations have been recently increased rapidly with respect to the depletion of fossil-based fuels all over the world. Due to stochastic nature of meteorological conditions, wind and solar energies have a non-schedulable nature and they require several installation analyses to determine the location and the capacities of wind and solar power to be produced.

This paper focuses on the similarity, feasibility and numerical analyses of 75 cities in Turkey based on the monthly average wind speed and insolation period data. The nearest and the farest neighbor algorithms are used as agglomerative hierarchical clustering methods with Euclidean, Manhattan and Minkowski distance metrics in the stage of making the similarity and feasibility analyses. The maximum cophenetic correlation coefficient is achieved by the nearest neighbor algorithm with the Minkowski distance metric in the similarity and feasibility analyses. On the other hand, graphical representations of the monthly average wind speed and insolation period data are utilized for making the numerical analysis. The highest annual average wind speed and insolation period are obtained as 3.88 m/s and 8.45 h/day, respectively. Overall, many inferences were achieved in acceptable and efficient limits for wind and solar energy.


## 1. Introduction

Wind and solar energies are inexhaustible and promising renewable energy resources for future. The utilization ratio of renewable energy sources in electricity generation has been expanded remarkably last couple of years in the world. The most crucial indicator of this case is that renewable energy based electricity generation was 3900 TWh in 2009 and it is expected as 11100 TWh in 2035 [1]. To reach that aim, installing new wind and solar energy stations become more crucial. The global cumulative installed wind power and solar power capacities increased to 198 GW and 40 GW in 2010, while wind power and solar power capacities are 6.1 GW and 0.7 GW in 1996 [2]. Turkey also has significant wind and solar energy potential because of its geographical characteristics. At the end of 2009, the total installed wind power and solar power capacities of Turkey have reached to 727.45 MW and 1 MW, respectively [3,4].

The annual average wind speeds were recorded as 4.29, 7.78, 7.4 and 9.98 m/s and the corresponding annual average wind power densities were quantified as 94.11, 570, 420 and 1177.97 W/m$^2$ in [5–8], respectively. The annual average wind power densities of Sani and Barisal were also estimated to be 452.5 and 94.94 W/m$^2$, respectively [9,10]. Samandağ and Amasra in Turkey were classified as nearly good in terms of the annual average wind power density by achieving the powers of 143 W/m$^2$ and 232 W/m$^2$ at the height of 10 m [11]. The wind power density for each month in the Cold Season exceeded 100 W/m$^2$ and the highest wind potential was appeared as 188 W/m$^2$ in February for Guelph in Canada [12]. It is shown that the wind potential in Germany has not allowed a significant reduction of fossil capacities and the cost-saving potential for electricity production has exceeded the subsidies [13].

The photovoltaic electricity generation in Taiwan had the potential of 36.1 TWh accounting for 16.3% of the total domestic electricity consumption [14]. The daily average solar radiations of Gaize and Florianópolis were ranged from 14 to 27 MJ/m$^2$ and from 2.46 to 5.72 kWh/m$^2$, respectively [15,16]. The annual average solar radiation was obtained between 750 kWh/m$^2$ and 2485 kWh/m$^2$ for Lake Van Region in Turkey [17]. The rooftop solar photovoltaic potentials were quantified as 885.1 kWh and 10 TWh in [18,19], respectively. The hybrid photovoltaic/thermal solar system increased the production of electrical and thermal energy by 38% [20]. The cumulative installed photovoltaic power increased to 1.45 GW in USA, 2 GW in Japan and 5.3 GW in Germany by the end of 2006–2008, respectively [21].

* Corresponding author. Tel.: +90 3122028538; fax: +90 3122120059.
*E-mail addresses:* icolak@gazi.edu.tr (I. Colak), ss@gazi.edu.tr (S. Sagiroglu), mehmetd@gazi.edu.tr (M. Demirtas), myesilbudak@nevsehir.edu.tr (M. Yesilbudak).

0196-8904/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.enconman.2012.07.011

**Nomenclature**

| | | | |
|---|---|---|---|
| m | meter | TWh | terawatt-hour |
| h | hour | $W/m^2$ | watt per square meter |
| m/s | meter per second | $kWh/m^2$ | kilowatt-hour per square meter |
| h/day | hour per day | $MJ/m^2$ | megajoule per square meter |
| $km^2$ | kilometer squared | GW | gigawatt |
| hPa | hectopascal | MW | megawatt |
| $kg/m^2$ | kilogram per square meter | NE | the angle between 33.75° and 56.25° |
| kWh | kilowatt-hour | NNE | the angle between 11.25° and 33.75° |

Most of the studies given above generally utilize Weibull model, Rayleigh distribution model, maximum entropy theory, artificial neural networks, feature extraction algorithms and geographical information systems in the stage of analyzing wind and solar power potentials. The agglomerative hierarchical clustering methods also have a wide range of different applications as in [22–25], but the usage of them is limited in the field of renewable energy sources. The aim of this paper is, on the one hand, to make the similarity, feasibility and numerical analyses of the monthly average wind speed and insolation period data belong to 75 cities in Turkey using agglomerative hierarchical clustering methods. On the other hand, to uncover the effects of the employed clustering approach and the selected distance metric on the similarity and feasibility analyses. As a result of this study, many reasonable, applicable and effective inferences have been achieved by clustering the monthly average wind speed and insolation period data hierarchically.

In the following of this article, Section 2 focuses on explanations of data mining and cluster analysis. Section 3 introduces UML (unified modeling language) activity diagram of the model developed. Section 4 explains interface layer of the model developed. Section 5 demonstrates the numerical analyses of the monthly average wind speed and insolation period data. Section 6 expresses the clustering analysis of the monthly average wind speed and insolation period data. The work is finally concluded in Section 7.

## 2. Data mining and cluster analysis

Data mining is expressed as a process of knowledge extraction used for revealing previously unknown, hidden, meaningful and useful patterns in databases [26]. It also covers an interdisciplinary structure including machine learning, pattern recognition, statistics, database management systems, intelligent systems and data visualization [27]. It is widely used in marketing, banking, finance, biology, chemistry, medicine, image recognition, telecommunications, geographic information systems, meteorology, astronomy, social sciences, behavioral sciences, power system management, energy trading, text mining and web mining [28]. Data mining techniques are categorized as characterization and discrimination, classification, cluster analysis, association analysis, outlier analysis and evolution analysis [26,29].

Cluster analysis has a wide range of application in the field of data mining and based on grouping data structures according to their similarities or differences [30]. Hierarchical methods, partitioning methods, density-based methods, grid-based methods and heuristic methods are used as clustering techniques [26]. In these clustering techniques, Euclidean, Manhattan and Minkowski distance metrics are generally used for calculating the distances between observations. The formulations related to these distance metrics are given in Eqs. (1)–(3) by assuming $p$ as the number of variables, $i,j = 1,2,\ldots,n$ and $k = 1,2,\ldots,p$ [31]. In Minkowski distance metric, Manhattan and Euclidean distance metrics are obtained by setting $m$ to 1 and 2, respectively. The metrics are given as:

$$\text{Euclidean distance metric}: d(i,j) = \sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2} \quad (1)$$

$$\text{Manhattan distance metric}: d(i,j) = \sum_{k=1}^{p}(|X_{ik} - X_{jk}|) \quad (2)$$

$$\text{Minkowski distance metric}: d(i,j) = \left[\sum_{k=1}^{p}(|X_{ik} - X_{jk}|^m)\right]^{\frac{1}{m}} \quad (3)$$

Hierarchical clustering methods are classified as the agglomerative and the divisive hierarchical clustering methods [32]. In the agglomerative hierarchical clustering methods, each observation in the dataset is evaluated as an independent cluster and a series of merge operations are carried out hierarchically until obtaining the cluster which contains all of observation. However, in the divisive hierarchical clustering methods, the entire dataset is considered as a cluster and a series of diverge operations are realized hierarchically until obtaining all of observations as singleton clusters [33,34].

The nearest neighbor (single-linkage) and the farest neighbor (complete-linkage) algorithms are used as agglomerative hierarchical clustering methods. In the nearest neighbor and the farest neighbor algorithms, initially, the distances between all of observations in dataset are calculated by means of a distance metric. Afterwards, the observations which have the highest similarity degree create a new cluster [35]. However, in case of merging the observation clusters, the distances between observations that are the nearest and the farest to each other among the observation clusters are assigned as the similarity degrees in the nearest and the farest neighbor algorithms, respectively [36]. These steps are repeated until the number of clusters is equal to one.

In this study, the nearest neighbor and the farest neighbor algorithms are used as clustering techniques and Euclidean, Manhattan and Minkowski metrics are used as distance metrics for the purpose analysis using the monthly average wind speed and insolation period of 75 cities in Turkey.

## 3. UML activity diagram of the model developed

UML is used for specifying, visualizing, constructing and documenting the artifacts of software systems and categorized as structure, action and interaction diagrams [37]. In this study, the activity diagram which is one of the action diagrams is utilized for presenting the model developed as shown in Fig. 1.

First of all, a user selects the clustering algorithm and the distance metric to be used in clustering operation. The nearest and the farest neighbor algorithms are used as agglomerative hierarchical clustering methods in the developed model. Euclidean, Manhattan and Minkowski distance metrics are also used for calculating the similarity degree in clustering operation. Afterwards, the dataset in Excel format containing monthly average