European Association for the Development of Renewable Energies,
Environment and Power Quality (EA4EPQ)

International Conference on Renewable Energies and Power Quality
(ICREPQ'10)
Granada (Spain), 23th to 25th March, 2010

Statistical Processing of Wind Speed Data for Energy Forecast and Planning

Annalisa Di Piazza[1], Maria Carmela Di Piazza[2], *Member IEEE*, Antonella Ragusa[2], *Member IEEE,*
Gianpaolo Vitale[2], *Member IEEE*

[1]Università degli Studi di Palermo
Dipartimento di Ingegneria Idraulica e Applicazioni Ambientali (DIIAA)
viale delle Scienze 90128 – PALERMO, ITALY
dipiazza@idra.unipa.it

[2]Consiglio Nazionale delle Ricerche
Istituto di Studi sui Sistemi Intelligenti per l'Automazione (ISSIA) sezione di Palermo
Via Dante, 12 90141 PALERMO, ITALY
TEL. +39 091 6113513 FAX +39 091 6113028
mariacarmela.dipiazza@ieee.org, ragusa@pa.issia.cnr.it , gianpaolo.vitale@ieee.org

Abstract. This paper presents a statistical approach to manage wind speed sampled data in order to obtain the forecast of the wind energy potential of a given site. The proposed statistical method is the k-means clustering that allows to extract from a set of experimental measurements the sub-sets of useful data for describing the energy capability of the site. The wind speed distributions in different sites in Sicily, in the south of Italy, have been studied as case study. A suitable wind generator, matching the wind profile of the studied sites, has been selected for the evaluation of the producible energy. It is demonstrated that the use of the proposed method simplifies the problem of the wind plant energy assessment respect to the option of obtaining the desired information by managing a large amount of experimental observations. The proposed method represents a useful tool for an appropriate energy planning in distributed generation systems.

Keywords

Wind energy; Distributed generation; Planning and control of the power systems; Statistical processing of data.

1. Introduction

The energy planning in a distributed generation system requires the appropriate knowledge of the renewable energy source capability. When a wind plant is involved, in particular, the energy capability is strictly correlated with the characteristic of the wind generator in terms of power curve. The power curve of a wind turbine shows the profile of the instataneous electric power versus the wind speed. In order to obtain the maximum efficiency of the wind generator, the wind speed of the considered site should maintain the value corresponding to the nominal power, as long as possible. Consequently, a study of the spatial and temporal wind distribution is needed for choosing or designing the generator that maximizes the wind energy potential of a given region.

The development of analytical tools for the estimation of the quantity of electrical energy generated by a wind plant on a given scale of time is therefore reputed very useful. In particular the possibility to find out, from the hystorical wind experimental observations, the most significant data to characterise the site of installation from the energy capability point of view is particularly advantageous.

The development of forecasting models for spatial and temporal distributions of climatic variables has been widely treated in technical literature, within the scope of energy assessment. In such a field the synergic use of suitable data processing techniques and estimation methods, either based on statistical or neural approach, represents the more promising way for the set-up of complete and reliable climatic databases and for the modelling and forecasting of the considered phenomena [1]-[3]. In this paper a statistical approach, based on the k-means clustering technique, is proposed in order to obtain an effective estimation of the energy produced by a wind plant in a one year, hourly sampled, scale of time [4]-[7]. The method is applied to the instantaneous power data and different sites in Sicily are explored, as for the wind speed profiles.

The proposed method allows to evaluate the contributions to the total annual energy production given by different ranges of the registered wind speeds. It allows also to eliminate the wind speed data which contribute in a marginal way to the amount of the produced energy. These data correspond to the lowest wind speeds and, even if numerous, they can be neglected and treated as outliers. In this way the energy forecast can be performed in a simpler way, by managing a reduced sub-set of data.

2. Wind Speed statistical distribution

The analytical representation of the wind speed probability distribution is useful when a problem of spatial projection or energy forecast has to be solved.

Fig.1 shows a typical distribution of the wind speed during one year. By observing fig.1 one can assess that the most probable wind speed value is lower than the average one. Therefore a Gaussian model of the probability distribution is not applicable. In general the wind speed is described by a two parameters Weibull distribution, whose probability density function (pdf) is given by:

$$f(U)=\frac{k}{C}\left(\frac{U}{C}\right)^{k-1}\exp\left[-\left(\frac{U}{C}\right)^{k}\right] \tag{1}$$

where k is the shape parameter, C the scale parameter and U the wind speed. Fig. 2 shows the pdf profile superimposed to the observed data [6]-[12].

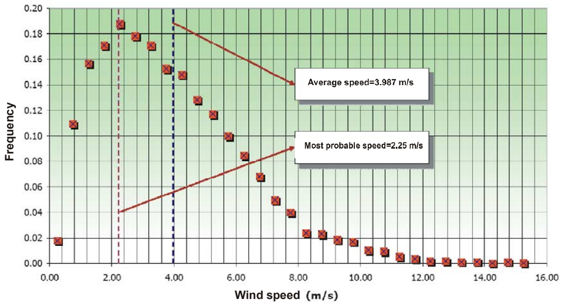

fig.1. Distribution of the wind speed during one year.

The cumulative distribution function of the Weibull distribution is given by:

$$F(U)=\exp\left[-\left(\frac{U}{C}\right)^{k}\right] \tag{2}$$

By multipliying this quantity for the cumulative number of hours in a year, the duration curve is obtained. This curve allows to evaluate the number of the hours in a year when a given speed value is exceeded.

The shape parameter is extremely important since it gives a direct information on the site typology; for example it reaches values of about 1.6 in mountain areas, values of about 2 in coastal areas and reaches values up to 3 in regions subjected to stationary or periodic winds.

For a given average speed, a lower value of k implies a greater available energy, as shown in fig. 3. It occurs because, with a low shape factor, the wind speed range is greater. In particular, since the generated energy depends on the cube of the speed, the contribution of low frequent strong winds to the total energy is significant. It is important to outline that the possibility to span the total wind speed values range depends on the operating range of the chosen wind turbine.

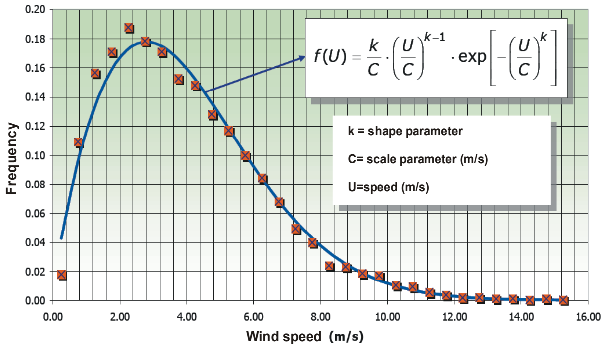

fig. 2. Two parameters Weibull pdf profile.

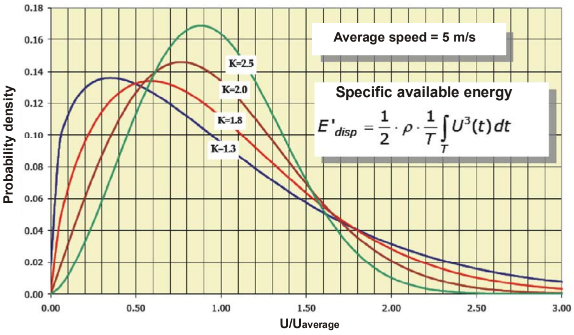

fig. 3. Influence of the shape factor, k, on the Weibull pdf.

The two parameters of the Weibull distribution can be calculated by a regression from the experimental data.

On the other hand, a simpler method for the identification of the parameters k and C can be used: starting from the wind speed values and their cumulative frequencies, two auxiliary variables (x and y) are defined according to the following relations:

$$x=\ln(U) \tag{3}$$

$$y=\ln[-\ln(1-F)] \tag{4}$$

The diagram that shows y versus x is described by the following straight line equation:

$$y=y_o+m\,x \tag{5}$$

The angular coefficient m is equal to k, while the term y_o allows to evaluate the scale factor C by the relation:

$$C=e^{-\frac{y_0}{m}} \tag{6}$$

Furthermore, if a maximum likelihood method is used, the two Weibull pdf parameters are obtained by the following equations:

$$\hat{k}=\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n}x_i^{\hat{C}}\right]^{\frac{1}{\hat{C}}} \tag{7}$$

$$\hat{C}=\frac{n}{\left(\frac{1}{n}\right)\sum_{i=1}^{n}x_i^{\hat{C}}\log x_i-\sum_{i=1}^{n}x_i} \tag{8}$$

being

$$\{x_i\}_{i=1}^{n}$$

the set of the n experimental data.

Once the parameters of the Weibull pdf are calculated, the mean and the variance are obtained.

The mean is given by:

https://doi.org/10.24084/repqj08.680

$$\mu_U = \mathrm{E}(U) = C\Gamma\left(1+\frac{1}{k}\right) \tag{9}$$

where

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$$

is the gamma function.
The variance is given by:

$$\sigma_U^2 = \mu_U^2 \left[\frac{\Gamma(1+2/k)}{\Gamma^2(1+1/k)} - 1\right] \tag{10}$$

In this paper the Weibull distributions describing the wind characteristics of the chosen sites have been identified, using some Matlab® routines. In particular the function *WBLFIT* returns maximum likelihood estimates of the parameters of the Weibull distribution given the data, the *WBLSTAT* function returns the mean and variance of the Weibull distribution for given shape and scale parameters and finally the *WBLPLOT* command displays a Weibull probability plot of the data. It should be noted that the purpose of a Weibull probability plot is to graphically assess whether the observed data could come from a Weibull distribution. If the data are Weibull-distributed, the plot is linear. As an example the Weibull pdf and the Weibull probability plot for the site of Pachino in the south east of Sicily are shown in figs. 4 and 5, respectively.

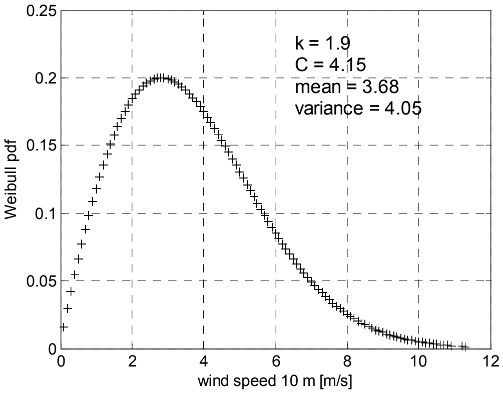

fig. 4. Weibull pdf for the average 10 m wind speed distribution in Pachino (one year, data sampled each hour).

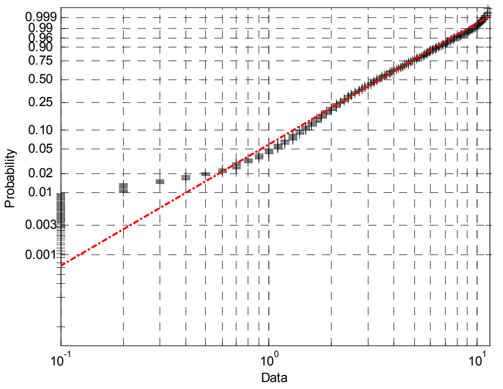

Fig. 5. Weibull probability plot for the average 10 m wind speed distribution in Pachino (one year, data sampled each hour).

3. Wind Generator Selection

The choice of the wind generator is done on the basis of the wind profile of the desired site. The aim is to obtain the maximum capacity factor for the installed generator, i.e. its maximum actual efficiency. This efficiency is measured in terms of equivalent duration (in hours) of the wind turbine operation at its nominal power. In fig. 6 the 10 m average wind speed distribution in some European countries is shown. It can be observed that the range of speed values in Sicily is between 3.5 and 5 m/s. The records of 10 m wind speed data of several sites in Sicily in one year have been considered, where the data are sampled each hour. On the basis of the observed data distributions, the 20kW wind generator VERGNET GEV 10/20 has been chosen. The generator has the features summarized in Table I [13].

Table I. Wind generator specifications

Wind generator model	Vergnet GEV 10/20
Rated power	20 kW
Cut-in wind speed	4.5 m/s
Rated wind speed	16 m/s
Maximal wind speed	60 m/s
Rotor diameter	10 m
Tower height	30 m

Since the wind speed has to be considered at the generator height, the observed data at 10 m are redefined at the height of 30 m by means of the following relation:

$$U(z) = U(z_{rif})\left(\frac{z}{z_{rif}}\right)^m \tag{11}$$

where z_{rif} is the reference height (10 m), z is the new height where the wind speed is evaluated and m is the roughness index calculated by the Counihan equation, and assuming values between 1 and 1.15, in Sicily.[14]-[15].

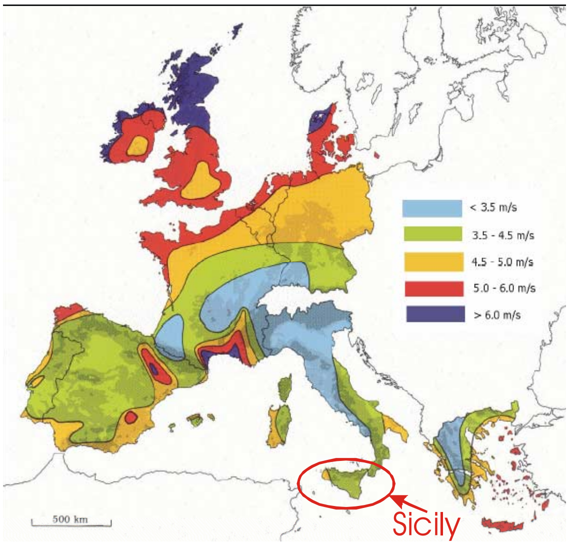

Fig.6. Wind speed chart for some European countries (average speed at 10 m).

Fig. 7 shows the power curve of the chosen generator [13]. It should be noted that the power curve of a wind generator represents a "medium" operating condition. It is obtained from a mathematical representation of the experimental data of power and wind speed, according to the "bin method" (described in the standard IEC 61400-12) [16]. These data are actually distributed in a cloud due to the extreme variability of the wind-related phenomena (turbulence, for example).

4. The k-means Clustering Method

The k-means clustering is basically a partitioning method. For a given set of observed data, the k-means method performs the partition of them into k mutually exclusive clusters.

Unlike the hierarchical clustering methods, k-means does not create a tree structure to describe the groupings in data, but rather creates a single level of clusters, using the actual observations of objects or individuals in data, and not just their proximities. These features make k-means more suitable for clustering large amounts of data, as in the case under study.

The k-means treats each observation in data as an object having a location in space.

It finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. Each cluster in the partition is defined by its member objects and by its centroid, or center. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized according to an assigned algorithm. The method of k-means computes clusters centroids, to minimize the sum with respect to a specified measure [3]-[6].

The application of k-means method for the partition of data coming from the studied wind plant has been performed within Matlab® environment. In particular the embedded *kmeans* function is used to obtain a vector of indices, indicating to which of the k clusters it has assigned each observation in data. Then, an algorithm, set-up on purpose by the authors, is employed to extract the sets of data assigned to each cluster.

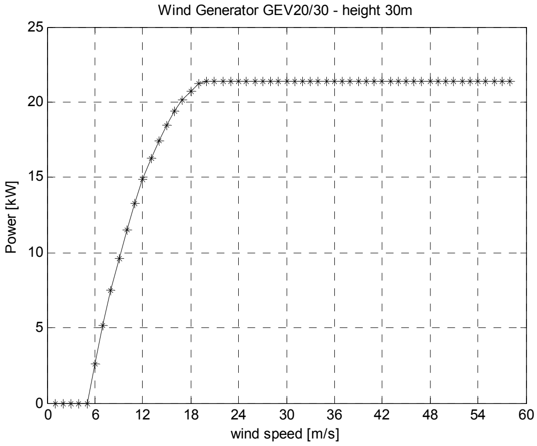

Fig.7. Power curve of the chosen wind generator [data sheet].

The *kmeans* function uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until the sum cannot be further decreased. The result is a set of clusters that are as compact and well-separated as possible.

To get an idea of how well-separated the resulting clusters are, it is useful to make a silhouette plot using the cluster indices output from *kmeans* function. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters. This measure ranges from +1, indicating points that are very distant from neighbouring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster.

In the described data processing the determination of the correct number of clusters is an issue. A possible way to compare the considered solutions is to look at the average silhouette values for different choices of number of clusters. In general it is a good idea to experiment with a range of values for k, selecting the value of k corresponding to a partition containing clusters with points having mostly high silhouette values. In this work the described method is used to select the appropriate number of clusters.

5. Total Energy calculation

In general the available wind energy, E_A in given site and for a given time period, T (one year, for example), is dependent on the cube of the wind speed u, the square of the turbine diameter and on the air density ρ, according to the following equation:

$$E_A = \int_0^T P_A \, dt = \int_0^T \frac{1}{2} \rho S u^3 dt \qquad (11)$$

where

$$P_A = \frac{1}{2} u^2 \left(\rho S u \right) = \frac{1}{2} \rho S u^3$$

is the kinetic power due to a given air mass flow passing through a section S.

In this paper the total electric energy produced by the wind generator in a time period T is calculated starting from the experimental observations by the relation:

$$E = \int_0^T P \, dt \qquad (12)$$

where P is the instantaneous electric power.

To be more precise, as the data acquisition is discrete (one acquisition each hour), the used energy expression is:

$$E = \sum_{i=1}^{n} P_i \cdot \Delta t \qquad (13)$$

The discrete values of the power P_i have been calculated starting from the wind speed data registrations and defining a function, *P=f(u)*, which fits the chosen generator power curve.

The obtained total energy is a cumulative value given by the sum of the energy contributions in defined time intervals. Such time intervals are deduced on the basis of the wind speed frequency distribution diagram. As an example, in figs. 8 and 9 show the 30 m wind speed frequency distribution diagrams, related to the site of Pachino and Trapani, respectively, for observations taken for one year. It can be observed that the total wind speed range, represented on the x-axis, has been divided into intervals having a width of 1 m/s, while the y-axis gives the cumulative duration of the different wind speeds in the considered period of one year.

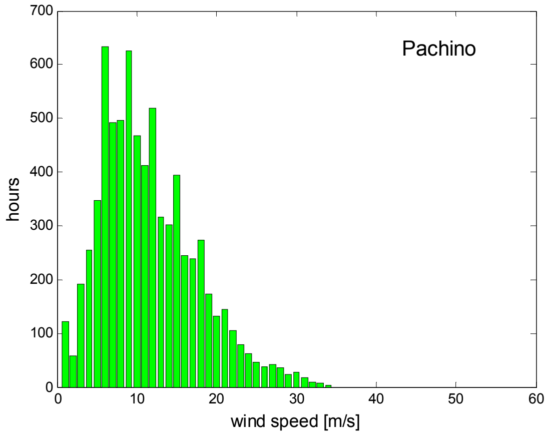

Fig.8. Wind speed frequency distribution diagram in Pachino (one year, data sampled each hour).

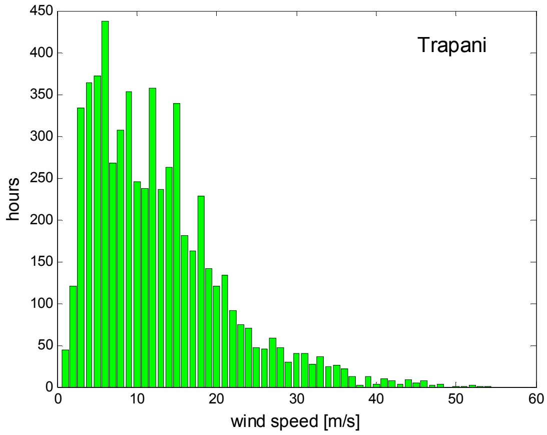

Fig.9. Wind speed frequency distribution diagram in Trapani (one year, data sampled each hour).

The total electric energy in one year, with a sampling of wind speed data performed every hour, has been calculated for the sites of Pachino (south east of Sicily) and Trapani (north west of Sicily), which have been chosen as case study.

The following results are obtained:

1. Total energy, for Pachino, $E_{(Pachino)}$= 85529 kWh;
2. Total energy, for Trapani, $E_{(Trapani)}$= 70345 kWh.

6. Energy assessment obtained through k-means clustering

A k-means-based partition of the overall observed wind speed data is performed in order to evaluate the presence of data sub-sets which allow to describe accurately the energy production of the wind plant. If it is the case, other data sub-sets can be neglected in the computation of the energy, since they can be regarded as outliers. For the scope, all the couples of observed wind speed and corresponding power in the chosen sites, in the period of observation of one year, have been taken as starting data set.

No definite knowledge of how many clusters are really in the data, by the energy point of view, is available. On the other hand, considering the shape of the generator power curve, the choice for *k* has been made on the basis of some experiments, for *k* ranging from 3 to 4. In particular the average value *h* of silhouette plots obtained for each value of *k* has been evaluated. The best clustering of starting data set has been obtained for *k*=3. In figs. 10 and 11 the silhouette plots corresponding to *k*=3 for the sites of Pachino and Trapani are reported, respectively. Both plots are obtained minimizing the sum of squared Euclidean distances from centroid for each cluster.

The number n_i (i=1 ÷ 3) of couple speed-power contained in each cluster is indicated. Extracting the data from each cluster, by means of an algorithm set-up by the authors, each contribution to the total wind generator energy production is evaluated.

The obtained results can be summarized as follows.

For Pachino the cluster 1, formed of 1990 couple of data, corresponding to the higher wind speeds, contributes to the total energy with an amount equal to $E_{(Pachino)\_1}$= 40760 kWh, i.e. the 48 % of the total energy; the cluster 2 formed of 2587 couple of data, corresponding to the medium wind speeds, contributes to the total energy with an amount equal to $E_{(Pachino)\_2}$= 35388 kWh, i.e. the 41% of the total energy; finally the cluster 3 formed of 2752 couple of data, corresponding to the lower wind speeds, contributes to the total energy with an amount equal to $E_{(Pachino)\_3}$= 9381 kWh, i.e. only the 11% of the total energy.

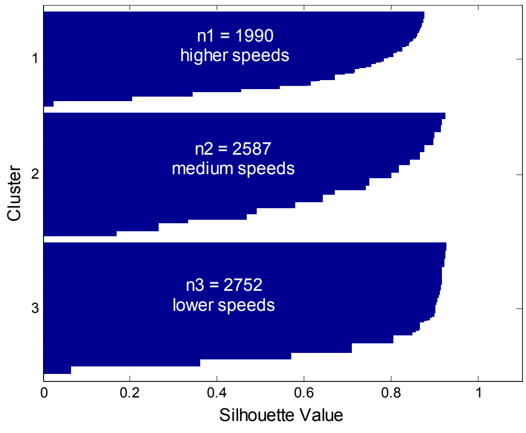

fig. 10. Silhouette plot for Pachino, k=3.

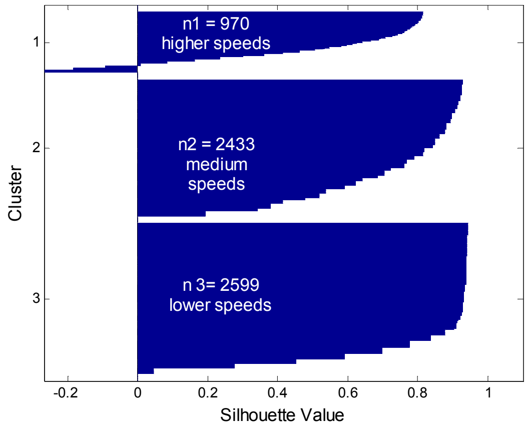

fig. 10. Silhouette plot for Trapani, k=3.

For Trapani the cluster 1, formed of 970 couple of data, corresponding to the higher wind speeds, contributes to the total energy with an amount equal to $E_{(Trapani)\_1}$= 20771 kWh, i.e. the 29 % of the total energy; the cluster 2 formed of 2433 couple of data, corresponding to the medium wind speeds, contributes to the total energy with an amount equal to $E_{(Trapani)\_2}$= 41392 kWh, i.e. the 59 % of the total energy; finally the cluster 3 formed of 2599 couple of data, corresponding to the lower wind speeds, contributes to the total energy with an amount equal to $E_{(Trapani)\_3}$= 8182 kWh, i.e. only the 12 % of the total energy. Table II summarizes the results of the energy evaluation obtained by clustering in comparison with the total energy, calculated in section 5. It is possible to observe that in both the considered sites, the main contribution to the total energy producible by the wind generator comes from data corresponding to the higher and medium wind speed independently from their number. This is due to the dependence of the energy on the cube of the wind speed, stated in eq. (11). Therefore the obtained results show that, with the proposed technique, a data sub-set containing negligible points can be identified. Then the energy capability for a given site can be computed, with a good approximation, considered only the data contained in the remaining clusters. On the basis of the given results it is possible to assess that the proposed method is a valid tool to characterize a site by the point of view of the energy capability coming from wind, managing only a reduced sub-set of data corresponding to the most significant ones.

7. Conclusions

A statistical approach based on the k-means clustering used to manage the wind speed data for the evaluation of the wind energy potential of a given site is proposed in this paper. The method makes possible the extraction from a set of experimental measurements of the sub-sets of useful data for describing the energy capability of the site. The wind speed distributions in one year in two sites in Sicily have been studied as case study.

A wind generator, matching the wind profile of the studied sites, has been selected for the evaluation of the producible energy. The obtained results show that the use of the proposed method allows the wind plant energy assessment using a reduced sub-set of experimental data, avoiding the management of a large amount of experimental observations.

Table II. Computed Energy

Site	Pachino	Trapani
Total energy [kWh] – CASE 1	85529	70345
Energy without the lower speed data contibution [kWh] – CASE 2	76148	62163
Percentage of deviation %	11	12
Managed data in CASE 1	7329	6002
Managed data in CASE 2	4577	3403

Acknowledgements

The authors thank the SIAS (Servizio Informativo Agrometeorologico Siciliano) - Regione Siciliana, for the delivering of the wind speed measured data.

References

[1] Jeffrey, Stephen J.; Carter, John O.; Moodie, Keith B.; Beswick, Alan R.; 2001 - Using spatial interpolation to construct a comprehensive archive of Australian climate data, in Environmental Modelling & Software, 2001, vol. 16, 309-330.

[2] Tang, W. Y.; Kassim A.H.M.; Abubakar, S. H.; 1996 - Comparative studies of various missing data treatment methods – Malaysian experience, in Atmospheric Research, 1996, vol. 42, 247-262.

[3] A. Di Piazza, M. C. Di Piazza, G. Vitale, “Statistical Processing of Data Coming from a Photovoltaic Plant for Accurate Energy Planning”, International Conference on Renewable Energy and Power Quality 2008 (ICREPQ08).

[4] Erto, P., “Probabilità e Statistica per le Scienze e l’Ingegneria”, McGraw-Hill, 2004.

[5] Davies, D.L., Bouldin, D.W., "A Cluster Separation Measure", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, 1979, pp. 224-227.

[6] Seber, G.A.F., “Multivariate Observations”, Wiley, New York, 1984.

[7] Seguro, J.V., Lambert, T.W., “Modern estimation of the parameters of the Weibull wind speed distribution for wind energy analysis”, J Wind Eng. Ind. Aerodyn 2000, 85(1):75-84.

[8] Ramirez P.,Carta J.A., “Influence of the data sampling interval in the estimation of the parameters of the Weibull wind speed probability density distribution: a case study”, Energy Convers. Manage, 2005, 46: 2419-38.

[9] Jain, A.K., Dubes, R.C., "Algorithms for Clustering Data", Prentice Hall, 1988, pp. 96-101.

[10] Celik A.N. “A statistical analysis of wind power density based on the Weibull and Rayleigh models at the southern region of Turkey”, Renew. Energy 2003, 29(4): 593-604.

[11]Weisser D.A., “Wind energy analysis of Grenada: an estimation using the ‘Weibull ’density function”, Renew. Energy 2003, 28(11): 1803-12.

[12]Lun I.Y.F.,Lam J.C.A., “Study of Weibull parameters using long-term wind observations”, Renew. Energy 2000, 20(2): 145-53.

[13]GEV 10/20 EN online available: http://www.sopac.org

[14] Counihan, J., “Adiabatic Atmospheric Boundary Layers: A Review and Analysis of Data Collected from the Period 1880-1972”, Atmospheric Environment, 1975, 9, 871-905.

[15] Rapporto Annuale del DPS – 2003, Note metodologiche – Cartine.

[16] IEC 61400-12: Wind turbine generator systems – Part 12: Wind turbine power performance testing.