Renewable Energy Generation and Load Classification based on H-K compound clustering algorithm

Yong Sun¹, Zhenyuan Li¹, Dexin Li¹, Chang Liu¹, Shengyan Wang²*, Fengkai Qiu² and Ming Zeng²

¹ State grid Jilin electric power co. LTD, Changchun, China
² School of economics and management, North China Electric Power University, Beijing, China

*Corresponding author e-mail: 1257177550@qq.com

Abstract. In this paper, the uncertainty of diesel generator (hereinafter referred to as DG) output and load demand is studied by multi scenario analysis. Firstly, this paper studies the timing characteristics of DG power generation and load demand, and describes the annual and daily distribution of photovoltaic power generation, wind power generation and load demand. Then it introduces the method of multi scene analysis, describes the process of dealing with the uncertainty in multi scene in detail, and generates enough scenes through the density function of classification probability to get the probability of each scene. Finally, H-K compound clustering algorithm is used to solve the above problems. The scale scene is compressed to get a typical "planning scene".

1. Overview of DG output and load demand uncertainty
With the large-scale DG access to the distribution network, the grid connected operation of DG has become an inevitable trend in the future power grid development. However, the common DG mainly uses wind and solar energy as the main primary energy [1]. The main feature of this type of energy is that it is greatly affected by external environmental factors, which leads to great uncertainty in the output of wind power generation system and photovoltaic power generation system. Therefore, research on grid investment optimization under the new situation has important theoretical and practical guiding significance for ensuring reasonable investment scale, precise direction, structural optimization, scientific timing, strict control of inefficient investment, and elimination of invalid investment.

2. Time series characteristics of DG output and load demand
The wind speed data and solar radiation intensity of a certain area can be obtained by Homer software, and the annual change curve of wind turbine, photovoltaic power generation system and various load demands can be obtained. By comparing and analysing the variation curves of DG output and load demand, the output value and load demand value of wind turbines and photovoltaic power generation systems are determined to be highly correlated with the seasons, so as to provide a basis for the later multi-scenario analysis to divide the time sequence scenarios in spring, summer, autumn and winter.
2.1. Timing of wind power output
The main influencing factor of wind power output is the wind speed at the wind turbine hub, which is greatly affected by the external factors such as the latitude and longitude of the wind turbine installation area and the seasonal climate. The wind speed change rule in spring, summer, autumn and winter is quite obvious, among which the mean value of wind speed from December to February and winter of the next year is the largest, followed by that in spring and autumn, but generally the wind speed in autumn is larger than that in spring, and the extreme value and mean value of wind speed in summer are the smallest [2]. The wind speed varies greatly in a day in this area. Generally, the wind speed increases slowly from about 10 a.m. and begins to decline around 18 p.m., but it is not absolute. The wind speed varies randomly.

2.2. Timing of photovoltaic output
The output power of photovoltaic power supply is greatly affected by the radiation intensity of sunlight, which often has intermittent and large uncertainty. Therefore, the photovoltaic output power will change due to the change of light intensity, which also shows randomness. The solar radiation intensity value is mainly concentrated from 6:00 a.m. to 18:00 p.m., and its distribution curve can be approximately regarded as obeying the normal distribution. In summer, the solar radiation intensity is the largest, and the average daily sunshine time is the longest, while in winter, the difference is obvious. In winter, the solar radiation intensity is the smallest, and the average daily sunshine time is the shortest. The change of solar radiation is also the change of photovoltaic output.

2.3. Complementary characteristics of wind power output
Based on the above analysis, the output value of wind turbine and solar photovoltaic power generation in summer and winter has typical seasonality [3]. Therefore, this paper further verifies whether the wind turbine output has complementary characteristics by selecting the output curve of wind turbine and solar photovoltaic power generation in summer and winter as an example. Therefore, in the planning of DG access distribution network, it is necessary to consider the timing characteristics of DG output and load demand.

3. Classification probability multi scenario analysis method

3.1. Overview of multi scenario analysis
This paper will consider the timing characteristics of the random variables in Section 3.2, and combine with the above scenario analysis method to form a multi scenario analysis method, so as to solve the uncertainty of the output and load demand of the distributed power supply, and apply it to the DG planning problem.

3.2. Scene generation based on classification probability
Firstly, the historical data of wind speed, light intensity and load demand are statistically analyzed, and the typical 24-hour data of each random variable in spring, summer, autumn and winter are obtained. Assuming that the error between the typical data and the actual data at each time is subject to the probability density function of its statistical characteristics, the expression of random variables in each "Planning scenario" is shown in formula (1).

\[
\begin{align*}
    v_{t,s} &= v^0_t + \Delta v_{t,s} \\
    E_{t,s} &= E^0_t + \Delta E_{t,s} \\
    L_{t,s} &= L^0_t + \Delta L_{t,s}
\end{align*}
\]  

(1)

Where \(v^0_t, E^0_t, L^0_t\) are wind speed value of wind power at time \(t\) of typical day, light intensity value and load demand value of photovoltaic; \(v_{t,s}, E_{t,s}, L_{t,s}\) are wind speed value of wind power at time \(t\)
of s scenario, light intensity value and load demand value of photovoltaic; $\Delta v_{t,s}$, $\Delta E_{t,s}$, $\Delta L_{t,s}$ are wind speed error value of wind power at time t of s scenario, light intensity error value and load demand error value of photovoltaic; $s = 1, 2, \cdots, N_s$ is for a period of Look at the total number of scenes; $t = 1, 2, \cdots, 24$.

Taking wind speed error as an example, the detailed steps are as follows.
Step 1: generate simulation data by Homer software to sort out wind speed data of 24 typical days in spring, summer, autumn and winter, as shown in table 1.

| Season | Spring | Summer | Autumn | Winter |
|--------|--------|--------|--------|--------|
|        | 4.5    | 3.5    | 5.1    | 6.1    |
|        | 5.1    | 3.6    | 5.5    | 6.5    |
|        | 4.9    | 3.7    | 5.9    | 7.1    |
|        | 4.8    | 3.9    | 6.5    | 7.5    |
|        | 4.7    | 3.9    | 6.3    | 7.3    |
|        | 5.1    | 3.7    | 6.2    | 6.9    |
|        | 5.5    | 3.9    | 6.1    | 6.7    |
|        | 6      | 4.1    | 7      | 7.3    |
|        | 7      | 4.2    | 8      | 8.5    |
|        | 7.1    | 4.3    | 9      | 9.2    |
|        | 7.7    | 5      | 9.5    | 10.1   |
|        | 13.4   | 11.9   | 15.3   | 17.3   |
|        | 13.9   | 12.1   | 16.1   | 18.5   |
|        | 14.5   | 12.5   | 16.9   | 19.9   |
|        | 14.9   | 13.5   | 17.5   | 20.7   |
|        | 15.1   | 13.1   | 18.5   | 21.6   |
|        | 14.9   | 12     | 18.1   | 19.7   |
|        | 14.5   | 11.5   | 17.5   | 18.3   |
|        | 5.1    | 4.3    | 6.5    | 9.9    |
|        | 5      | 4.2    | 6.3    | 9.1    |
|        | 4.7    | 4.1    | 6.1    | 8.2    |
|        | 4.6    | 3.8    | 5.9    | 7.3    |
|        | 4.5    | 3.7    | 5.5    | 6.9    |
|        | 4.6    | 3.6    | 5      | 6.5    |

Step 2: Assuming that the wind speed error $\Delta v$ is a random variable, it can be approximately considered to meet the Weibull distribution characteristics, and its probability density function can generally be expressed as:

$$f(\Delta v) = \exp\left[ -\left(\frac{\Delta v}{c}\right)^k \right] \frac{\Delta v}{c} \left(\frac{\Delta v}{c}\right)^{k-1}$$

(2)

In the formula, $c$ and $k$ are the shape parameters and scale parameters of Weibull distribution respectively, which can be obtained from the mean value of the errors of historical wind speed and typical daily wind speed at each time and the standard deviation 6. The calculation formula is shown in formula (3) and formula (4).

$$k = \left(\frac{\sigma}{\mu}\right)^{-0.086}$$

(3)

$$c = \frac{\mu}{\Gamma(1+k^{-1})}$$

(4)
Where, $\Gamma$ is gamma function. Taking the wind speed error at 12:00 noon as an example, the parameter values of the four seasons are obtained from formula (3) and formula (4), as shown in Table 2.

Step 3: take the probability density function that the wind speed error obeys at 12 noon in spring as an example, and divide it into 7 intervals as shown in Figure 1. The abscissa is the wind speed error, the width of each interval $\sigma_{\Delta}$ is 4, and the ordinate $\alpha_{x,t}(x = 1, 2, \cdots, 7)$ is the probability density corresponding to each interval, that is, the probability corresponding to each interval is $\alpha_{x,t} \cdot \sigma_{\Delta}$. The probability density function that the wind speed error obeys at 12 noon in spring is an example, and divide it into 7 intervals as shown in Figure 1. The abscissa is the wind speed error, the width of each interval $\sigma_{\Delta}$ is 4, and the ordinate $\alpha_{x,t}(x = 1, 2, \cdots, 7)$ is the probability density corresponding to each interval, that is, the probability corresponding to each interval is $\alpha_{x,t} \cdot \sigma_{\Delta}$.

Table 2. Weibull distribution parameters of typical daily wind speed in four seasons

|       | Spring | Summer | Autumn | Winter |
|-------|--------|--------|--------|--------|
| $k$   | 1.96   | 1.71   | 1.97   | 2.11   |
| $c$   | 9.29   | 7.16   | 10.83  | 12.56  |

Figure 1. Dispersion of the PDF of speed error

The sum of probabilities for all intervals as shown in Figure.

In the same way, for the error values of light intensity and load demand, they are approximately subject to Beta distribution [4] and normal distribution [5] respectively through statistical analysis. The steps of solving the probability density function are similar to those of wind speed error, which is limited to the space problem, and will not be discussed here.

Step 4: the roulette method [6] is used to determine the selected interval of each random variable error in each scene, which is represented by a set of binary numbers. A number between 0~1 is randomly generated, and the selected interval is set to 1, and the rest interval is set to 0, so as to determine the occurrence probability of each random variable in each scene, as shown in formula 5.

$$s = \{\Delta v_{x,t,s}, \Delta v_{y,t,s}, \Delta v_{z,t,s}, \Delta E_{x,t,s}, \Delta E_{y,t,s}, \Delta E_{z,t,s}, \Delta L_{D,x,t,s}, \Delta L_{D,y,t,s}, \Delta L_{D,z,t,s}\}$$  (5)

In the formula, $\Delta v_{x,t,s}, \Delta E_{y,t,s}, \Delta L_{D,x,t,s}$ respectively represents the selection of wind speed error range, light intensity error range and load power error range at time $t$ in the scene; D represents the load type ($D = 1, 2, 3)$.

Step 5: calculate the probability of each scene, as shown in equation (6).

$$\omega(s) = \prod_{i=1}^{24} \left( \sum_{x=1}^{2} \left( \Delta v_{x,t,s}, \beta_{x,t,s}, \sigma_{\Delta}\right) \sum_{y=1}^{2} \left( \Delta E_{y,t,s}, \beta_{y,t,s}, \sigma_{\Delta}\right) \prod_{D=1}^{3} \left( \sum_{z=1}^{2} \left( \Delta L_{D,x,t,s}, \gamma_{z,t,s}, \sigma_{\Delta}\right) \right) \right)$$  (6)

In the formula, $\omega$ is the probability of occurrence of $s$ in each scene, $\alpha_{x,t} \cdot \sigma_{\Delta}$, $\beta_{y,t} \cdot \sigma_{\Delta}$, $\gamma_{z,t} \cdot \sigma_{\Delta}$ is expressed, respectively, the probability of occurrence of wind speed error interval at time $t$, illumination intensity error interval and load demand power error interval.

$$P_t(s) = \frac{\omega(s)}{\sum_{s=1}^{N} \omega(s)}$$  (7)
In the formula, $P_i$ is the normalized probability for each scenario is given.

3.3. Scene reduction based on H-K composite clustering

In this paper, density-based K-means clustering method [7] and Hierarchical aggregation clustering algorithm (HAC) are chosen to fuse to form H-K composite clustering algorithm. Take $n$ objects to be clustered as an example, set the initial number of center scene clusters to 3, and the cluster center is represented by a star. The clustering process is shown in Figure 15. Among them, figure a is the space formed by the collection of all the scenes to be allocated; figure B is the three randomly selected scenes in this space as the centroid scene; figure C is the state formed after the first allocation of all the scenes to the initial centroid scene through the principle of maximum density and shortest distance; figure D is the state of the new cluster center generated by the first adjustment of the old cluster center according to the average method; figure e is the state of the new cluster center generated by the second iteration via The density maximum distance minimum principle updates the state graph of each cluster. After many adjustments, if the new and old cluster centers change within a certain allowable error range, the whole iteration process will end and the final cluster and cluster centers clustering effect will be formed as shown in Fig. 2.

Assuming that NS "planning scenes" are generated in the scene generation phase, the number of scenes obtained after HAC clustering is $t_s$, and the number of typical "planning scenes" expected to be obtained is $M_s$, the specific steps of scene reduction using H-K composite clustering algorithm are as follows:

1) firstly, HAC algorithm is used. At the beginning, each scene is regarded as an initial class, with $N_s$ class in common, and the minimum distance between the two is calculated:

$$d_{min}(k_i, k_j) = \min_{p \in k_i, p' \in k_j} \| p - p' \|_2$$

$k_i = 1, 2, \cdots, N_s; k_j = 1, 2, \cdots, N_s;$

2) Find the closest two classes and merge them into a new class. At this time, the total number of classes is $N_s - 1$;

3) recalculate the distance between the newly generated class and all the old classes;

4) Repeat steps 2) to 3) until the total class number is $t_s$;

5) Take the TS clusters obtained by HAC algorithm as the input of K-means algorithm, randomly select $M_s$ scenes as the center of the cluster, and the set of cluster center scenes is $Center = \{ T_s^{Center} \} (s = 1, \cdots, M_s).$

6) From the set of cluster-centered scenes, it can be concluded that the set of the remaining scenes at this time is $Member = \{ T_s^{Member} \} (s' = 1, \cdots, T_s - M_s).$ Calculate the distance from the remaining scene to the central scene of each cluster:

$$d(T_s^{Center}, T_s^{Member}) = \| T_s^{Center} - T_s^{Member} \|_2$$
7) According to the distance matrix $d_{s,s'}$, the remaining scenes to be assigned are assigned to the nearest cluster center. At the end of the first clustering operation, the cluster set is $\text{Sort} = \{S_i\}_{i=1}^M$, where $S_i$ represents a cluster with high similarity of internal scene.

8) Selection method of typical "planning scene" cluster center: assuming that there are $LS$ scenes in a cluster $S_i$, calculate the sum of distance between each scene and other scenes respectively:

$$ds = \sum_{s' \neq s}^{L_s} \|T_s - T_{s'}\|, s = 1, 2, \cdots, L_s$$

The selected scene $TK$ is the new center of the cluster $d_k = \min(d_s)$. In the same way, the cluster center scene set of all clusters is determined.

9) Repeat 6) to 8) steps until the clustering results and the typical "planning scenarios" obtained are no longer changed, and the whole process of scene reduction is completed. The probability value of the typical "planning scene" of each cluster center is equal to the sum of the probabilities of all the scenes in the scene set.

4. Summary of this chapter
Firstly, this paper gives a brief overview of the uncertainty of DG output and load demand, and illustrates its necessity in the study of DG planning. Secondly, the timing characteristics of DG output and load demand are studied, and the annual and daily distribution of PV and wind power output and load demand are described with representative pictures. Then, the multi scenario analysis method is introduced.

Acknowledgments
This work was financially supported by State Grid Technology Project (522300190009): Research and application of the key technologies of source-demand participation in market transactions based on data value discovery.

References
[1] Pan W X, Zhou J, Wang F. A scenario technology applied to wind speed simulation in wind farm [C]. International Conference on Power System Technology. IEEE, 2014:2775-2780.
[2] Zhi-Gang L U, Sui Y S, Shou-Long H E, et al. Economic/unit ramp dispatching for power grid integrated with wind power based on scenario analysis [J]. Advanced Technology of Electrical Engineering&Energy, 2016.
[3] Jamshid Aghaei, Taher Niknam, Rasoul Azizipanah Abarghooe, et al. Scenario-based dynamic economic emission dispatch considering load and wind power uncertainties [J]. Electrical Power and Energy Systems, 2013(47): 351-367.
[4] Zhong Z, Yang C, Cao W, et al. Short-Term Photovoltaic Power Generation Forecasting Based on Multivariable Grey Theory Model with Parameter Optimization [J]. Mathematical Problems in Engineering, 2017(7): 1-9.
[5] Ming Z, Yang Y, Xiang H, et al. Optimal dispatch model based on coordination between"generation-grid- load- energy storage" and demand-side resource [J]. Electric Power Automation Equipment, 2016. Kohl L, Bruce D, Croft D. Roulette Gaming Method, System And Apparatus., US20160328905 [P]. 2016.
[6] Kohl L, Bruce D, Croft D. Roulette Gaming Method, System And Apparatus: US20160328905 [P]. 2016.
[7] Zhang X, Xu Z. Hesitant fuzzy agglomerative hierarchical clustering algorithms [J]. International Journal of Systems Science, 2015, 46(3): 562-576.