**Clustering Wind Speed Data by Hyperbolic Smoothing Clustering Method**

Vinicius Layter Xavier*
Federal University of Rio de Janeiro, Brazil – viniciuslx@gmail.com

José Francisco Moreira Pessanha
Rio de Janeiro State University - UERJ, Brazil – professorjfmp@hotmail.com

Luiz Antônio Alves de Oliveira
Electric Power Research Center - Eletrobras Cepel, Rio de Janeiro, Brazil – laao@cepel.br

Adilson Elias Xavier
Federal University of Rio de Janeiro, Brazil – adilson@cos.ufrj.br

**Abstract**

The wind is an abundant and clean resource to electric power generation, but it is a non controllable resource. The wind variability put challenges to the integration of wind power plants to the electric grid. In this context, the understanding about daily profile of wind speed can mitigate the effects of the uncertainties in wind behaviour on the electric power system. The typical daily profiles of wind speed are valuable information to short-term wind speed forecast and to the design of wind power plants. The typical daily profiles can be obtained by clustering techniques applied to a set of wind speed measurements. In this paper we show the typical profiles of wind speed at an anemometric station operated by the SONDA project. The typical profiles were obtained by Hyperbolic Smoothing Clustering Method (HSCM), a relatively new and powerful method for cluster analysis.

**Keywords:** wind speed; wind power; cluster analysis; hyperbolic smoothing clustering method.

## 1. Introduction

Wind is a renewable natural resource and its utilization in the electricity production is one of the most promising alternatives to mitigate air pollution and to replace fossil fuels in the electric power generation.
The statistics from World Wind Energy Association (WWEA, 2014) show that the total worldwide installed wind capacity reached 336 GW by mid-2014, a growth of 13.5% over mid-2013. Estimates indicate that the worldwide installed wind capacity by mid-2014 generated around 4 % of the world's electricity demand. However the rapid growth experienced by wind power points out to the increasing of its share in the total electric power production. These perspectives are based on fact that wind energy is a matured technology and its development will increase its competitiveness relative to other sources of electricity.
Although in Brazil the electricity production is predominantly hydroelectric, therefore clean and renewable, recently the country became the third largest market for new wind turbines, with 1.3 GW of new capacity (WWEA, 2014). The wind power potential in Brazil is estimated on 143.5 GW at 50 m height (AMARANTE et al., 2001). The harnessing of this potential can complement the hydroelectric generation during droughts periods and increase the power supply reliability.
The wind power production depends on the wind speed, a random variable. In order to cope with the stochastic behavior of the wind, the integration of wind farms into the electrical power systems requires short-term forecasts of wind speed. In this context the knowledge of the typical wind speed daily profiles contributes to the wind power forecasts (KIMI et al, 2011) and therefore for the management of the wind farms (DUARTE et al, 2012). In addition, the typical wind speed profiles are basic information to the design of wind turbines (GÓMEZ-MUÑOZ & PORTA-GÁNDARA, 2002;

AZEVEDO et al, 2014) and to the filtering of wind speed measurements in order to correct outliers and fill missing data (PESSANHA et al, 2011).
The typical daily wind speed profiles can be obtained by clustering techniques applied to a sample of wind speed measurements organized in a data matrix which each line represents a day and each column represents a time of day. Briefly, a cluster analysis method classifies objects into groups such that similar objects are classified in the same group. The typical profiles are obtained by averaging the daily profiles assigned to the same cluster.
Among the clustering methods available, the Hyperbolic Smoothing Clustering Method – HSCM (XAVIER, 2010; XAVIER & XAVIER, 2011) has emerged as one of the most promising algorithms for cluster analysis. The HSCM considers the solution of the minimum sum-of-squares clustering - MSSC problem. The mathematical modeling of this problem leads to a min-sum-min formulation which has the significant characteristic of being strongly non-differentiable.
The proposed method adopts the Hyperbolic Smoothing (HS) strategy using a special $C^{\infty}$ differentiable class function. The final solution is obtained by solving a sequence of low dimension differentiable unconstrained optimization sub-problems which gradually approach the original problem. The proposed algorithm applies also a partition of the set of observations into two non overlapping groups: "data in frontier" and "data in gravitational regions". The resulting combination of the HS methodology with the partition scheme for the MSSC problem has interesting properties, which drastically simplify the computational tasks.
Therefore, this study aims to investigate the use of hyperbolic penalty algorithm to identify the typical daily wind speed profile. In order to show the performance of the HSCM on this clustering problem, the paper presents the results from computational experiments with a four-year wind speed measurements recorded at 10-min intervals from anemometric station at São João do Cariri – Brazil, sited at latitude 7º22'54" S and longitude 36º31'38" W with an altitude of 486 m, operated by the National Network of Environmental Data for Renewable Energy Resource Assessment (SONDA project, http://sonda.ccst.inpe.br/). Next, in section 2 we present the fundamental smoothing procedures, in section 3 we describe the Hyperbolic Smoothing Clustering Method. The results from computational experiments are resumed in section 4 and in section 5 we show the main conclusions.

## 2. The Fundamental Smoothing Procedures

The core idea of the approach is the smoothing of the clustering problem formulation by a smoothing scheme, called Hyperbolic Smoothing (HS), presented in Santos (1997) for non-differentiable problems in general. This technique was developed through an adaptation of the hyperbolic penalty method originally introduced by Xavier (1982) in order to solve the general non-linear programming problem.
By smoothing it fundamentally mean the substitution of an intrinsically non-differentiable two-level problem by a $C^{\infty}$ differentiable single-level alternative. This is achieved through the solution of a sequence of differentiable sub-problems which gradually approaches the original problem. Each sub-problem, owing to its being indefinitely differentiable, can be comfortably solved by using the most powerful and efficient algorithms, such as conjugate gradient, quasi-Newton or Newton methods.
The methodology is based on the hyperbolic smoothing of the non-differentiable functions belonging to the optimization problem formulation. We will present the two basic smoothing procedures. First, we will consider the smoothing of the Euclidean distance between two generic points $s_j$ and $x_i$ belonging to $\Re^n$:

$$\theta(s_j, x_i, \gamma) = \sqrt{\sum_{l=1}^{n} \left(s_j^l - x_i^l\right)^2 + \gamma^2} \qquad (1)$$

Function $\theta$ has the following properties:

(a) $\lim_{\gamma \to 0} \theta(s_j, x_i, \gamma) = \left\| s_j - x_i \right\|_2$;

(b) $\theta$ is a $C^{\infty}$ function.

For smoothing the function $\varphi(y)=\max(0,y)$ we use:

$$\phi(y,\tau) = (y + \sqrt{y^2+\tau^2}\ )/2 \tag{2}$$

Function $\phi$ has the following properties:

(a) $\lim_{\tau\to 0}\phi(y,\tau) = \varphi(y)$;

(b) $\phi(y,\tau)$ is an increasing convex $C^\infty$ function in variable $y$.

### 3. Hyperbolic Smoothing Clustering Method

Let $S=\{s_1,s_2,...,s_m\}$ denote a set of $m$ observations in an Euclidean n-dimensional space $\Re^n$, to be clustered into a given number $q$ of disjoint clusters. To formulate the original clustering problem as a $\min-sum-\min$ problem, we proceed as follows. Let $x_i$, $i=1,...,q$ be the locations of centroids. The set of these centroids coordinates will be represented by $X \in \Re^{nq}$. Given a point $s_j \in S$, we initially calculate the Euclidian distance from $s_j$ to the nearest centroid: $z_j = \min_i \left\| s_j - x_i \right\|_2$.

The most frequent measurement of the quality of a clustering associated to a position of $q$ centroids is provided by the minimum sum of squares of these distances.

$$\text{minimize} \sum_{j=1}^{m} (z_j)^2 \tag{3}$$

$$\text{subject to:} \quad z_j = \min_i \left\| s_j - x_i \right\|_2, \qquad j = 1,\cdots,m.$$

By using function $\varphi$ and by performing a perturbation $\varepsilon>0$, we obtain the problem:

$$\text{minimize} \sum_{j=1}^{m} (z_j)^2 \tag{4}$$

$$\text{subject to:} \quad \sum_{i=1}^{q} \varphi(z_j - \left\| s_j - x_i \right\|\ ) \geq \varepsilon, \qquad j = 1,\cdots,m.$$

By using function $\phi$ in the place of function $\varphi$ and by using the approximation smooth function $\theta$ of the distance $\left\| s_j - x_i \right\|_2$ the problem turns into

$$\text{minimize} \sum_{j=1}^{m} (z_j)^2 \tag{5}$$

$$\text{subject to:} \quad h_j(z_j,x) = \sum_{i=1}^{q} \phi\left(z_j - \theta\left(s_j,x_i,\gamma\right),\tau\right) - \varepsilon = 0, \quad j = 1,\cdots,m$$

The dimension of the variable domain space of problem (5) is $(nq+m)$. As, in general, the value of the parameter $m$, the cardinality of the set $S$ of the observations, is large, problem (5) has a large number of variables. However, it has a separable structure because each variable $z_j$ appears only in one equality constraint. Therefore, as the partial derivative of $h_j(z_j, x)$ with respect to $z_j$, $j=1,\ldots,m$ is not equal to zero, it is possible to use the Implicit Function Theorem to calculate each component $z_j$, $j=1,\ldots,m$ as a function of the centroid variables $x_i$, $i=1,\ldots,q$. This way, the unconstrained problem:

$$\text{minimize} \ f(x) = \sum_{j=1}^{m} (z_j(x))^2 \tag{6}$$

is obtained, where each $z_j(x)$ results from the calculation of a zero of each equation

$$h_j(z_j,x) = \sum_{i=1}^{q} \phi\left(z_j - \theta\left(s_j,x_i,\gamma\right),\tau\right) - \varepsilon = 0, \quad j = 1,\cdots,m \tag{7}$$

Again, due to the Implicit Function Theorem, the functions $z_j(x)$ have all derivatives with respect to the variables $x_i$, $i$=1,…,$q$ and therefore it is possible to calculate the gradient of the objective function of problem (6)

$$\nabla f(x) = \sum_{j=1}^{m} 2\, z_j(x)\, \nabla z_j(x) \tag{8}$$

where

$$\nabla z_j(x) = -\nabla h_j(z_j, x) \Big/ \frac{\partial h_j(z_j, x)}{\partial z_j} \tag{9}$$

while $\nabla h_j(z_j, x)$ and $\partial\, h_j(z_j, x) / \partial\, z_j$ are obtained from equation (7) and from the definitions of functions $\varphi$ and $\theta$.

## 4. Computational Experiments

In order to evaluate the performance of the HSMC method we led a computational experiment with wind speed (m/s) recorded at 10-min intervals from anemometric station at São João do Cariri – Brazil. The measurements cover the period from January 1, 2006 to September 20, 2009. The dataset contains 1,327 complete wind speed daily profiles at 50 m height. The average wind speed in the period is 5.24 m/s. The boxplots of the average daily wind speed in Fig. 1 show that the lowest wind speeds are observed in April and May. In addition, Fig. 1 shows that the average wind speed increases during the dry season (between May and November), but it follows a downward trend in the wet season (between December and April). This is an example of the energy complementarity between wind and hydro sources in Brazil (PALFI & ZAMBON, 2013).

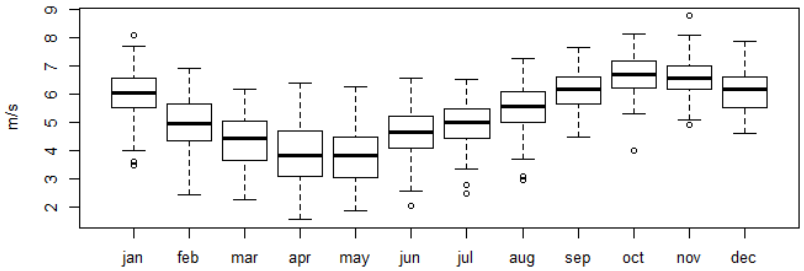


Figure 1. Boxplots of the average daily wind speed

The data matrix has dimension 1,327x144 and the Total Sum of Squares (TSS) is 0.6934161E$^6$. Below we present the results from computational experiment with the HSMC method in order to cluster the daily wind profiles. The numerical experiments have been carried out on a Intel Core i7-2620M Windows Notebook with 2.70GHz and 8 GB RAM. Table 1 presents a synthesis of the computational results. We vary the number of clusters from 2 to 12 and for each number of clusters ten different randomly chosen starting points were used. The columns show the number of clusters ($q$), the best solution produced (objective function $f_{best}$), the number of occurrences of the best solution (Occurrences), the average deviation of the 10 solutions ($E_{Mean}$) in relation to the best solution obtained, the ratio BSS/TSS (BSS Between Sum of Squares), the ratio WWS/TSS (WSS Within Sum of Squares), Pseudo-F (LATTIN et al, 2003), Compactness and Separation – CS (XIE & BENI, 1991) and CPU mean time given in seconds ($T_{Mean}$) associated to 10 tentative solutions.

The minimum value for CS statistic suggests four clusters while the maximum Pseudo-F suggests only two clusters, but in both solutions the ratio WSS/TSS concentrates more than 50% of TSS. Then, in order to explore the results we considered the solution with twelve clusters shown in Fig. 2, where we can see the daily wind profiles classified in each cluster and the respective centroid. In Fig. 2 we can observe some similar centroids. In fact Fig. 2 reveals that we can consider less than twelve clusters, perhaps a number between 2 or 4 clusters, in accord with the validity measures (Pseudo-F and CS).

| $q$ | $f_{best}$ | Occurrences | $E_{Mean}$ | BSS/TSS | WSS/TSS | Pseudo-F | CS | $T_{Mean}$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 0.476840E6 | 9 | 0.00 | 31.2 % | 68.8 % | 601.7 | 13.83 | 0.11 |
| 3 | 0.398418E6 | 1 | 0.00 | 42.5 % | 57.5 % | 490.1 | 14.92 | 0.21 |
| 4 | 0.352468E6 | 1 | 0.12 | 49.2 % | 50.8 % | 426.6 | 13.40 | 0.58 |
| 5 | 0.328345E6 | 8 | 0.02 | 52.6 % | 47.4 % | 367.5 | 14.34 | 0.46 |
| 6 | 0.314980E6 | 2 | 0.24 | 54.6 % | 45.4 % | 317.5 | 16.29 | 1.06 |
| 7 | 0.302397E6 | 1 | 0.49 | 56.4 % | 43.6 % | 284.5 | 15.11 | 2.82 |
| 8 | 0.291699E6 | 2 | 0.45 | 57.9 % | 42.1 % | 259.5 | 15.45 | 2.67 |
| 9 | 0.284448E6 | 1 | 0.64 | 59.0 % | 41.0 % | 236.9 | 16.64 | 3.99 |
| 10 | 0.278538E6 | 1 | 0.32 | 59.8 % | 40.2 % | 218.0 | 17.98 | 7.33 |
| 11 | 0.272940E6 | 1 | 0.53 | 60.6 % | 39.4 % | 202.7 | 17.88 | 7.71 |
| 12 | 0.268189E6 | 1 | 0.36 | 61.3 % | 38.7 % | 190.0 | 18.39 | 11.56 |

Table 1. Results from HSMC for different numbers of clusters

Fig. 3 shows a map obtained by correspondence analysis (GREENACRE, 2007) that reveals the associations between months and clusters. For example, the patterns in the clusters 1, 2, 3, 7 and 8 are more frequent in wet season while patterns in the clusters 6, 10, 11 and 12 are more frequent in the dry season. In Fig. 3, with the exception of May, the months in the dry season appear above the horizontal axis.

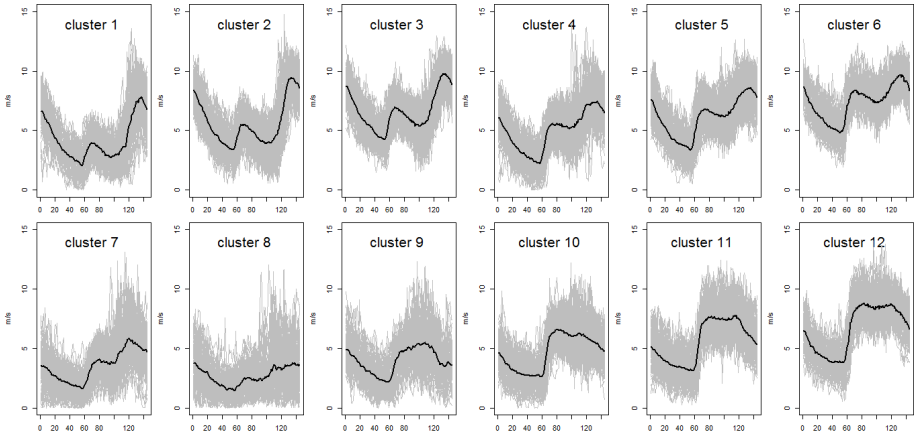


Figure 2. Clusters and centroids

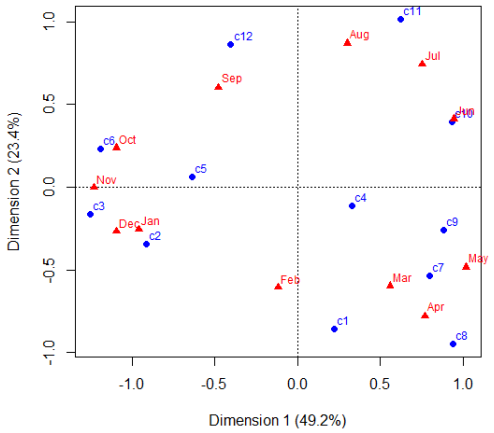


Figure 3. Map from correspondence analysis

## 5. Conclusions

We present results from HSMC method applied to the identification of typical daily wind speed profiles in a large dataset of wind speed measurements. The wind speed profiles are a valuable information to the development and operation of wind power projects. The robustness, consistency and efficiency performances achieved by the HSMC can be attributed to the complete differentiability of the approach.

## References


AMARANTE, O.A.C.; BROWER, M.; ZACK, J.; SÁ, A.L. Atlas do Potencial Eólico Brasileiro, Centro de Pesquisas de Energia Elétrica, Brasília, 2001.

AZEVEDO, P.A.A.; SOUZA, M.C.L.; PESSANHA, J.F.M.; BICALHO, J.R.S.; PECORELLI PERES, L.A. Uso da análise de agrupamentos na avaliação de aerogeradores de pequeno porte para certificação no programa brasileiro de etiquetagem. Acta Scientiae e technicae, v. 2, p. 37-45, 2014.

DUARTE, F.J.; DUARTE, J. M. M.; RAMOS, S.; FRED, A.; VALE, Z. Daily Wind Power Profiles Determination using Clustering Algorithms, IEEE International Conference on Power System Technology (POWERCON), Auckland, 2012.

GÓMEZ-MUÑOZ, V.M.; PORTA-GÁNDARA, M.A. Local wind patterns for modeling renewable energy systems by means of cluster analysis techniques, Renewable Energy, 25, pp. 171-183, Renewable Energy, 2002.

GREENACRE, M. Correspondence Analysis in Practice, 2nd ed., Chapman & Hall, 2007.

KIMI, K.I.; JIN, C.H.; LEE, Y.K.; KIM, K.D.; RYU, K.H. Forecasting wind power generation patterns based on SOM clustering, 3rd International Conference on Awareness Science and Technology (iCAST), Dalian, 2011.

LATTIN, J.; CARROLL, J.D.; GREEN, P.E. Analysing Multivariate Data, Thomson Learning, 2003.

PALFI, G.; ZAMBON, R. Hydro and Wind Power Complementarity and Scenarization in Brazil. World Environmental and Water Resources Congress, Cincinnati, 2013.

PESSANHA, J.F.M.; CASTELLANI, V.; JUSTINO, T.C.; JARDIM, D.L.D.D.; MACEIRA, M.E.P. A methodology for data cleaning of wind speed time series. X Brazilian Congress on Computational Intelligence (in Portuguese), Fortaleza, 2011.

SANTOS, A.B.A. Problemas de programação não-diferenciável: Uma metodologia de suavização. M.Sc. Thesis, COPPE-UFRJ, Rio de Janeiro, 1997.

XAVIER, A.E. Penalização hiperbólica: Um novo método para resolução de problemas de otimização, M. Sc. Thesis, COPPE - UFRJ, Rio de Janeiro, 1982.

XAVIER, A.E. The hyperbolic smoothing clustering method, Pattern Recognition, v. 43, pp. 731--737. 2010.

XAVIER, A.E.; XAVIER, V.L. Solving the minimum sum-of-squares clustering problem by hyperbolic smoothing and partition into boundary and gravitational regions, Pattern Recognition, v. 44, pp. 70-77, 2011.

XIE, X.L.; BENI, G.A. A validity measure for fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 13, n. 8, pp. 841–847, 1991.

WWEA - World Wind Energy Association, Half-year Report 2014, September, 2014. (http://www.wwindea.org/wwea-publishes-half-year-report-2014/).