

Identifying Representative Days of Wind Speed in Brazil Using Machine Learning Techniques

Rafaela Ribeiro

Pontifical Catholic University of Rio de Janeiro, Department of Industrial Engineering
Rua Marquês de São de Vicente, 225 – Gávea, Rio de Janeiro, RJ – Brazil – 22541-900
rafaelamaribeiro@aluno.puc-rio.br

Advisor: Bruno Fanzeres

Pontifical Catholic University of Rio de Janeiro, Department of Industrial Engineering
Rua Marquês de São de Vicente, 225 – Gávea, Rio de Janeiro, RJ – Brazil – 22541-900
bruno.santos@puc-rio.br

RESUMO

A geração eólica no Brasil tem atraído altos níveis de investimentos nos últimos anos devido a sua construção em curto prazo e com custo relativamente baixo, junto ao incentivo mundial para o estabelecimento de um sistema de energia neutro em carbono. Neste sentido, estudos sobre o comportamento e sobre a dinâmica característica futura de seu principal combustível (velocidade do vento) durante os dias e os anos provam-se extremamente necessários. Portanto, neste trabalho, serão usadas técnicas de *Machine Learning* para encontrar padrões similares da velocidade do vento em dados horários variantes no tempo, a fim de identificar dias representativos comuns, cruciais para planejamento, *designing*, operação e avaliação de sistemas de energia com fontes renováveis. São apresentadas análises descritivas baseadas em dados reais em locais com alta capacidade instalada de energia eólica atual e altos níveis potenciais de produção, mais especificamente, duas regiões brasileiras: Nordeste e Sul.

PALAVRAS CHAVE. Velocidade do Vento. Técnicas de Agrupamento. K-Medoids. Partitioning Around Medoids (PAM). Métodos de Agregação. Dias Representativos.

EN&PG – OR on Energy, Oil and Gas Areas, BDA – Big Data and Analytics

ABSTRACT

Wind power generation has attracted high investment levels in the past years in Brazil due to its relatively low construction costs and general short-time building construction, along with the worldwide incentive towards the establishment of a carbon-neutral power system. In this context, studies over the behaviour and future dynamic characteristics of its main fuel (wind speed) during the days and the years are proving to be extremely necessary. Therefore, in this work, we leverage on Machine Learning techniques to identify similar patterns on time-linked hourly-based wind speed data, aiming at identifying common representative days, crucial for planning, designing, operation, and assessment of renewable-based energy systems. A set of descriptive analytics is presented based on real data from locations with current high wind power installed capacity and potential production levels, namely, the two Brazilian regions: Northeast and South.

KEYWORDS. Wind Speed. Clustering Techniques. K-Medoids. Partitioning Around Medoids (PAM). Aggregation Method. Representative Days.

EN&PG – PO na Área de Energia, Petróleo e Gás, BDA – Big Data e Analytics

1. Introduction

Driven by the ambitious goals of a climate-neutral economy in the near future, renewable sources are gaining more and more space in the power generation mix worldwide. In Brazil, for instance, wind-based power production is already a significant source of electricity. In fact, it represented roughly 8.6% of the Brazilian total electricity supply in 2019, surpassed only by hydro and natural gas sources, which accounted for approximately 65% and 9% [EPE, 2020]. More specifically, in 2019, the Northeast Region of Brazil had a total wind power production of 47.1 TWh and the South Region an equivalent of 5.6 TWh, representing approximately 56% and 6% of the year total energy consumption in the respective region. Moreover, wind generation in Brazil increased 17% in 2019 over 2018, from 46.47 TWh to 54.37 TWh, leading the country to the seventh position on the GWEC (Global Wind Energy Council) rank of cumulative onshore wind capacity [Lee e Zhao, 2020]. In 2019, 38 new wind farms, spread over three states, were built in the country, all in Northeast Region: *Bahia* and *Rio Grande do Norte*, the two states that account for the highest levels of wind power generation, and *Maranhão*, considered the state with the largest average capacity factor in the country [ABE, 2020]. Also, ABE [2020] highlights the five Brazilian states with the highest power generation in 2019 as, from the highest to the lowest, *Bahia*, *Rio Grande do Norte*, *Piauí*, *Ceará* and *Rio Grande do Sul*.

In contrast to their environmentally-friendly operations, however, renewable sources are characterized by an intrinsic uncertainty in energy production [Moreira et al., 2018; Gielen et al., 2011]. Therefore, the development of techniques to aid regulatory agencies, central planners, and private companies to better comprehend this uncertain factor aiming at enhancing their capability to trade, plan, control, and operate power systems with large renewable penetration is of great importance, all in order to assure this affirmative growing pace of renewable sources into the grid [Passos et al., 2014; Moreira et al., 2019]. The methodology proposed in this work thus becomes particularly important in reducing dimensionality in the uncertainty representation for the planning and operations of the Brazilian electrical grid due to its singular geographic characteristics. In this context, the main objective of this work is to provide a descriptive analytics of the time-linked hourly-based wind speed uncertain behaviour over the years by making use of a methodology to group similar days. More specifically, the process of grouping is based on a Machine Learning clustering technique known as the *K-Medoids*, which has an unsupervised algorithm that clusters similar observations so that the distance between groups and the distance between the elements inside each group are minimized [Igual e Seguí, 2017; Bramer, 2016; Aggarwal, 2015].

The focus of our analysis is on wind farm sites located in the states of *Bahia*, *Rio Grande do Norte*, *Ceará* and *Rio Grande do Sul*, four of the five states with the highest wind power installed capacity in Brazil, with the latter belonging to the South Region and the remainder to the Northeast Region. On the one hand, we found a well-behaved hourly-based wind speed dynamics in the Northeast Region over the year, with cluster centres as decent representations and good candidates for dimensionality reduction. On the other hand, however, the hourly-based wind speed dynamics in the South Region do not possess a highlighted pattern (e.g., strong seasonality) over the year and all clusters are almost equally divided among the months. The main contribution of this work is therefore to provide a detailed description of the wind speed dynamics in different parts of Brazil so that future analysis (e.g., planning and operation) can be performed.

2. Methodology for Selection of Representative Days

Aiming for a better understanding of how renewable production uncertainty affects the electric power systems trading, planning, and operations, it is common adapt dimensionality reduction techniques to identify the so-called *representative days* (we refer to, for instance, Pinto et al.

[2020]; Hassan et al. [2021] for an extensive review and discussion). In this context, it is possible to make accurate decisions in a more effective way since this reduction enables the decision-maker to work in a lower dimension rather than the entire dataset that oftentimes can be hard to manipulate. In this work, to determine these representative days, we leverage on a Machine Learning clustering technique known as the K-Medoids [Dinov, 2018; Aggarwal e Reddy, 2013]. More specifically, this clustering technique is a partitioning and aggregation method for which is available an unsupervised algorithm that aim at separating the dataset into a number of K clusters so that the elements within each cluster are *similar* and outside the cluster are *different*. In other words, it minimizes the distance between observations to form the clusters' centre, hereinafter referred to as representative days for the context of this work. For most partitioning and aggregation methods, data normalization is typically recommended for computational purposes. This procedure should be done before the application of the method and is commonly used when the dataset encompasses time series from distinct origins, thus possess different units (see, for instance, the discussion in [Kotzur et al., 2018; Hoffmann et al., 2020]). In this work, we make use of the so-called min-max normalization [Hoffmann et al., 2020] to avoid numerical issues when applying the partitioning and aggregation method.

2.1. The K-Medoids Method

The K-Medoids method is an adaptation of the K-Means that, instead of calculating the average of the observations within each cluster to represent the centre, they are actual data observations [Schütz et al., 2018; Kotzur et al., 2018]. The K-Medoids is commonly used when the data is a set of network or a discrete sequence of objects due to the higher interpretability of its results and cluster centers, compared to, for instance, the K-Means, since it is often not meaningful to analyze functions (e.g., the average) of these objects [Aggarwal e Reddy, 2013]. Methodologically, this method can be formulated as a Mixed Integer Linear Programming (MILP) problem [Vinod, 1969] for which exists efficient computational procedures to handle it (e.g., the Partitioning Around Medoids (PAM) algorithm). More precisely, for an *a priori* defined number of K clusters, PAM follows the following steps [Scott et al., 2019]:

Step 1: Randomly select K days as centre of the clusters;

Step 2: Determine the Euclidian distance between each day and each K centres;

Step 3: Group each day with one of the centres following the least-distance (**Step 2**) criterion;

Step 4: Update the centres calculating, for every day, the distance between them and all other days. It should, as well, find what day minimizes the total distance to the other days that also belong to the cluster and determine it as its centre;

Step 5: Repeat **Steps 2–4** until the selected days to be centre do not change.

3. Descriptive Analytics

The wind speed data was selected so that it could represent four of the five states in Brazil with the highest wind power production. The considered wind farms were chosen to represent the sites with the highest annual average wind speed, following two main criteria: (i) it should be one of the wind farms with the highest energy production from the state it belongs; and (ii) it should be in operation. In this context, there were selected to study one wind farm from the states of *Ceará*, *Bahia* and *Rio Grande do Sul* and two from *Rio Grande do Norte*, as both operate individually but are located in the same site. With those choices, it is possible to adequately analyse the regions with the highest wind power production and, also, it is a good geographic representation of wind farm distribution in Brazil. All data was obtained on the public database: <https://www.renewables.ninja/> [Staffell e Pfenninger, 2016]. In this work, the wind speed data

was measured by the Vestas V90 2000 turbine model at a height of 80 meters, for the years since construction up to 2019. The information they provide come from the dataset MERRA-2 (global), from NASA [Staffell e Pfenninger, 2016]. To help understand the results and analysis presented in Section 4, in the following subsections, we thoroughly describe each region's characteristics and weather patterns.

3.1. South Region

The Macroscale influences the wind in the South Region as the heat between the Equator and the poles are not uniform. Furthermore, because of the Coriolis Effect (inertial force that explains the Kinetic Law, which, for the case of the Southern Hemisphere, drives the particles that move from the pole to the Equator towards the left), it also originates the geostrophic wind, a horizontal theoretical wind that flows in straight paths on the free atmosphere, situated on the top of the boundary layer. In the Mesoscale, the changes on wind speed happens mostly because of the topography and because of abrupt shifts of the superficial changes on the superficial reflection coefficient (typically referenced as albedo) which generates the katabatic wind (a common high density draught on cold regions that, due to the action of gravity, flows downwards on hillsides), most common on mountain regions and land breezes close to the littoral. As discussed in Wind-Atlas [2014], the *Rio Grande do Sul* state is located on a Transitional Climate Zone. This Zone is impacted both by intertropical systems, that are tropical air masses and phenomena as *El Niño* and *La Niña*, and by extratropical systems, i.e., masses and fronts that come from the South Pole and can cause South Atlantic Subtropical Anticyclones. On the Mesoscale, the most common climatological alterations are from the Northwestern Argentinean Low and The Chaco Low. When both connect with the South Atlantic Subtropical Anticyclone, the wind stays on a moderate speed. There is also the occasion of *Minuano* (or *Pampeiro*) Winds, that are cold winds that come from the Southeast with intensity superior to 10 m/s [Wind-Atlas, 2014]. Analysing the temporal scale, it is possible to notice that, in the littoral of *Rio Grande do Sul*, by the end of the afternoon (right before sunset) the air over the sand rapidly cools, making the air to flow from the land to the sea, characterizing thus land breezes. Moreover, in the littoral, from September to December, the wind speed is intensified, although with a modest level magnitude (see Fig. 4). As a consequence, based on the above-mentioned characteristics, we conclude that the wind in South Region presents fast variations of intensity and direction.

3.2. Northeast Region

The east coast and the semiarid of the Northeast Region of Brazil are the most influenced areas by the trade winds (i.e., strong and stable winds near to the Earth's surface) of which, close to the Equator, flows from east to west, because of the Coriolis Effect¹ [Torres Silva dos Santos e Santos e Silva, 2013]. The Intertropical Convergence Zone (ITCZ) occurs on the Equator and influences the interannuality variability on Northeast Region, mostly on the east coast and the semi-arid region. With respect to seasonality aspects, dry seasons (Winter and Spring) have higher wind speeds (and more atypical values), with the peak levels between August and October, and rainy seasons (Summer and Autumn – from February to May) have lower wind speeds. The Sea Breezes are responsible for the intensification of the trade winds during the dry seasons. The wind farms located in the east coast and the semiarid region (both wind farms from *Ceará* – Fig. 1, and *Rio Grande do Norte* – Fig. 2) have greater intensification and higher wind speed than the one on centre of Northeast Region (the one in the *Bahia* State – Fig. 3) due to the stronger sea breezes in the

¹We refer to National Oceanic and Atmospheric Administration – US Department of Commerce on <https://oceanservice.noaa.gov/facts/tradewinds.html>

littoral and weaker sea breeze in the interior. The *El Niño* is a phenomenon that induces the increase of Sea Surface Temperature (SST) of the Pacific Ocean and, according to Torres Silva dos Santos e Santos e Silva [2013], it is also responsible to less intense trade winds. On the other hand, the *La Niña* phenomenon is the opposite, causing the SST of the Pacific Ocean to decrease and, therefore, to increase the trade wind intensity.

4. Clustering Analysis

In this section, we present for all wind sites considered in this work the monthly distribution of the clusters along the year and the average wind speed of each month. In the first graphic, the clusters with the slowest wind speed are shown in cooler colours, the ones with the highest wind speeds are in warmer colours and in light yellow, that can be seen as a transition between cool and warm colours, the moderate wind speeds clusters. Notice, in the latter graphic, that the range of the wind speed varies for each wind farm. For expository purposes, we assume $K = 10$ clusters.

4.1. Praia Formosa Wind Farm, Camocim, Ceará (CE)

As presented in Fig. 1, *Praia Formosa* Wind Farm clearly has a pattern of which most clusters with smaller wind speed are located in the beginning of the year and the ones with higher wind speeds at the end. In fact, the warmer months (Summer and beginning of Autumn), from the beginning of the year, have smaller wind speeds, that rarely exceeds 10 m/s (clusters represented by the strongest shades of blue). In June, on the other hand, the average wind speed starts to increase (see Fig. 1), reaching higher average speed levels in September and, lastly, decreasing in November. This represents the increase of wind speed in Winter (end of June until end of September), the maintaining of it on the higher-level during Spring and the final decrease in the end of Spring (December). The transition is represented by the lighter blue and the shades of yellow, and the highest wind speeds by shades of orange.

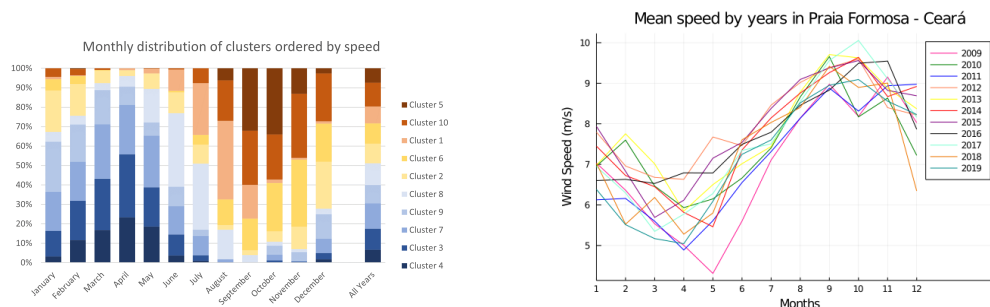


Figure 1: *Praia Formosa* – Clusters Monthly Distribution along the Year and Monthly Average Wind Speed

4.2. Alegria I and II Wind Farms, Guamaré, Rio Grande do Norte (RN)

As expected, the pattern on *Rio Grande do Norte* wind farms is, in general, similar to the one that occurs in *Ceará* (i.e., most of the clusters with lower wind speed levels are located in the beginning of the year and the ones with higher levels at the end) as observed in Fig. 2. In this context, in the beginning of the year, in the hottest months (Summer and beginning of Autumn), the wind speed is low (shades of blue). In June, on the other hand, the wind speed starts to increase, reaching higher levels in August, and it starts to decrease November. This represents an intensification on wind speed during the Winter (end of June to end of September) and a decrease during the Spring (November and December). The transition is shown by the clusters with shades of yellow and the highest wind speed levels by the shades of orange. Note that, for *Alegria I* and *Alegria*

II wind farms, Cluster #4 (the one that accounts for the smallest wind speed levels) does not have any element from August and November (months with the highest average wind speed levels). A similar pattern occurs on *Praia Formosa* wind farm (see Fig. 1). The remaining clusters with small wind speed (#3, #7, #10 and #9) have few days as well on this time of the year (Cluster #9 has the most days, followed by clusters #7, #10 and #3). The transition clusters appear on all months of the year and the ones with the highest wind speed levels have more days from July to December, even though they appear between January and June (mostly on the first two months of the year).

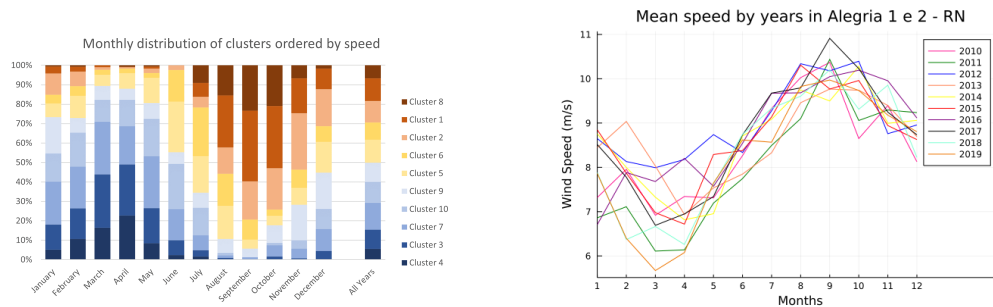


Figure 2: *Alegria I and II* – Clusters Monthly Distribution along the Year and Monthly Average Wind Speed

4.3. *Dos Araçás* Wind Farm, *Caetitê*, *Bahia* (BA)

The pattern in *Dos Araçás* Wind Farm is somewhat similar to its Northeast Region counterparts (*Ceará* and *Rio Grande do Norte*), but less delimited (see Fig. 3). Most clusters with lower wind speed levels are located in the beginning of the year and the ones with higher levels at the end. More precisely, on the hottest months (end of Spring, Summer and beginning of Autumn) the wind speed level is low, typically below 10 m/s, and is represented by the shades of blue. From June to October, on the other hand, the wind speed increases, changing its level. This represents an intensification during Winter and Spring. The transition is shown by clusters with shades of yellow and the highest speeds by shades of orange. It is possible to see that only two clusters do not have days from all months of the year. These clusters are the #1, that does not have days in February and March (expected since these months have lower speeds levels and Cluster #1 is characterized by higher ones), and the Cluster #4, which does not encompass days from June and August, for similar reasons. It worth highlighting that in the year of 2017 an increase above the typical in the monthly average speed was observed during the months of July and September (see Fig. 3). As discussed in Section 3.2, the *La Niña* phenomenon was the main responsible for this event.

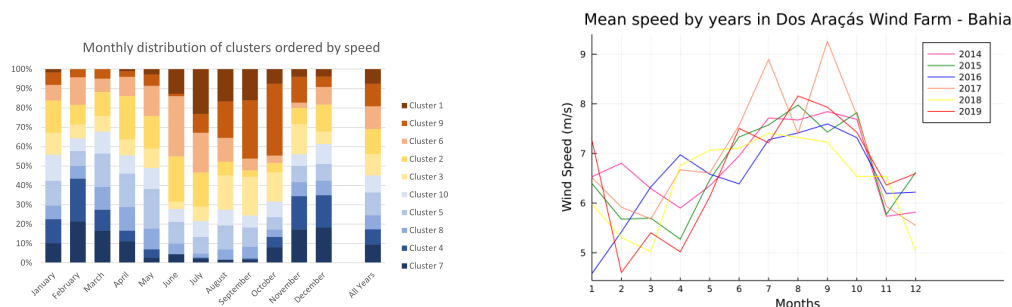


Figure 3: *Dos Araçás* – Clusters Monthly Distribution along the Year and Monthly Average Wind Speed

4.4. Elebrás Cidreira 1 Wind Farm, Tramandaí, Rio Grande do Sul (RS)

For the South Region, note that all clusters are evenly distributed along the year, i.e., encompass days that belong (uniformly) to all of the 12 months of the year (see Fig. 4). By analyzing the monthly average wind speed by year and the distribution of the clusters by month – Fig. 4, we note the nonexistence of any seasonal patterns in this region. Even though there might exist a slight increase in the average wind speed levels in July, this event is not sufficiently intense as the ones from the Northeast Region, thus the wind speed level is maintained until November (months of Winter and Spring).

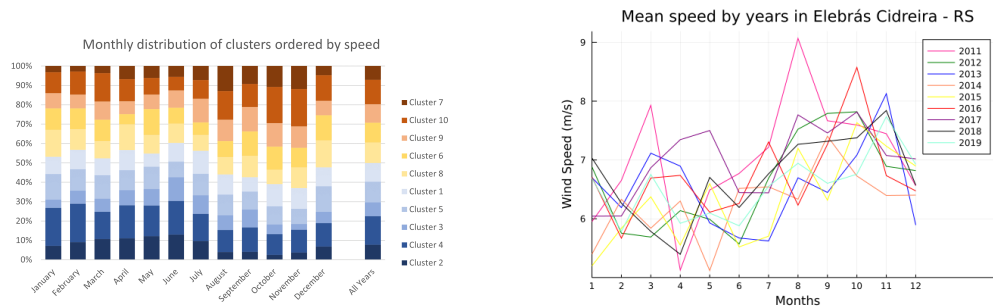


Figure 4: *Elebrás Cidreira 1* – Clusters Monthly Distribution along the Year and Monthly Average Wind Speed

5. Conclusion

Wind power generation is increasing and is already one of the most important electricity supply sources in Brazil. In this work, we analyse the time-linked, hourly-based wind speed dynamics over the year on sites with wind farms located in three states in the Northeast, the Brazilian region with highest wind power installed capacity, and also one located in the South, the Brazilian region with the second higher wind power installed capacity. The analysis was made using the Machine Learning clustering technique known as the K-Medoids method, in order to identify the so-called representative days of a year. In summary, for the South Region, we identified no clear pattern on the hourly-based wind speed over the years. On the other hand, nevertheless, it was found that the wind pattern in Northeast Region follows a seasonal pattern during the months and, moreover, that exists “standard days” as good candidates for dimensionality reduction, critical for efficient power system studies. Future work might focus on the analysis of the relationship between the wind speed clusters and electricity energy operation or generation, for example, rather than on weather. Furthermore, we also highlight that intraday, hourly patterns should be analyzed.

Bibliographical References

- (2020). Annual Wind Energy Report 2019. Technical report, ABEEólica.
- (2020). Brazilian Energy Balance – Summary Report 2020. Technical report, Energy Research Office – Brazil.
- Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.
- Aggarwal, C. C. e Reddy, C. K. (2013). *Data Clustering: Algorithms and Applications*. Chapman and Hall/CRC.
- Bramer, M. (2016). *Principles of Data Mining*. Springer, 3rd edition.

- Dinov, I. D. (2018). *Data Science and Predictive Analytics: Biomedical and Health Applications using R*. Springer, 1st edition.
- Gielen, D., Boshell, F., Saygin, D., Bazilian, M. D., Wagner, N., e Gorini, R. (2011). The Role of Renewable Energy in the Global Energy Transformation. *Renewable Energy*, 36(11):3068–3074.
- Hassan, M. A., Khalil, A., e Abubakr, M. (2021). Selection Methodology of Representative Meteorological days for Assessment of Renewable Energy Systems. *Renewable Energy*, 177:34–51.
- Hoffmann, M., Kotzur, L., Stolten, D., e Robinius, M. (2020). A Review on Time Series Aggregation Methods for Energy System Models. *Energies*, 13(3):641.
- Igal, L. e Seguí, S. (2017). *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*. Springer.
- Kotzur, L., Markewitz, P., Robinius, M., e Stolten, D. (2018). Impact of Different Time Series Aggregation Methods on Optimal Energy System Design. *Renewable Energy*, 117:474–487.
- Lee, J. e Zhao, F. (2020). Global Wind Report 2019. Technical report, Global Wind Energy Council.
- Moreira, A., Fanzeres, B., e Strbac, G. (2018). Energy and Reserve Scheduling under Ambiguity on Renewable Probability Distribution. *Electric Power Systems Research*, 160:205–218.
- Moreira, A., Strbac, G., e Fanzeres, B. (2019). An Ambiguity Averse Approach for Transmission Expansion Planning. In *2019 IEEE Milan PowerTech*, p. 1–6.
- Passos, A. C., Street, A., Fanzeres, B., e Bruno, S. (2014). A Novel Framework to Define the Premium for Investment in Complementary Renewable Projects. In *2014 Power Systems Computation Conference*, p. 1–7.
- Pinto, E. S., Serra, L. M., e Lázaro, A. (2020). Evaluation of Methods to Select Representative Days for the Optimization of Polygeneration Systems. *Renewable Energy*, 151:488–502.
- Schütz, T., Schraven, M. H., Fuchs, M., Remmen, P., e Müller, D. (2018). Comparison of Clustering Algorithms for the Selection of Typical Demand Days for Energy System Synthesis. *Renewable Energy*, 129:570–582.
- Scott, I. J., Carvalho, P. M. S., Botterud, A., e Silva, C. A. (2019). Clustering Representative Days for Power Systems Generation Expansion Planning: Capturing the Effects of Variable Renewables and Energy Storage. *Applied Energy*, 253:113603.
- Staffell, I. e Pfenninger, S. (2016). Using Bias-Corrected Reanalysis to Simulate Current and Future Wind Power Output. *Energy*, 114:1224–1239.
- Torres Silva dos Santos, A. e Santos e Silva, C. M. (2013). Seasonality, Interannual Variability, and Linear Tendency of Wind Speeds in the Northeast Brazil from 1986 to 2011. *The Scientific World Journal*.
- Vinod, H. D. (1969). Integer Programming and the Theory of Grouping. *Journal of the American Statistical Association*, 64(326):506–519.
- Wind-Atlas (2014). Wind Atlas: *Rio Grande do Sul*. Technical report, Camargo Schubert Engenheiros Associados, Eletrosul Centrais Elétricas S.A., <https://sema.rs.gov.br/atlas-eolico-do-rs>.