Dynamic Vector Model Applied to Wind Speed Prognosis for Eolic Generation

Aramis Perez[1], Francisco Cornejo[2], Marcos Orchard[3], and Jorge Silva[4]

[1,2,3,4]*Department of Electrical Engineering, Faculty of Physical and Mathematical Sciences, Universidad de Chile, Santiago, Chile*

aramis.perez@ing.uchile.cl
fcornejo@gmail.com
morchard@ing.uchile.cl
josilva@ing.uchile.cl

ABSTRACT

Dynamic characterization of energy availability profiles is paramount for an adequate incorporation of Non-Conventional Renewable Energies. This fact is particularly significant for sizing and design of eolic energy parks. The integration of eolic parks with interconnected systems requires accurate and precise knowledge on maximum and minimum power availability, as well as the moments in which you should expect the aforementioned conditions. Prognosis tools can help to determine the wind speed with a certain degree of reliability, in order to forecast energy availability. In this regard, this article aims at designing and implementing a methodology to generate a dynamic vector-autoregressive-based models for wind speed prognosis. This methodology makes use of techniques such as data clustering, time series statistical analysis and its characterization through time-variant parametric models, for a medium term horizon. The proposed method is able to prognosticate wind speed for a complete day in just one step, instead of classic approaches that repeat several one-step ahead transitions to obtain similar results. The employed methodology facilitates the identification of periodical components of the wind, including daily and seasonal, facilitating the differentiation of data clusters with similar behaviors or tendencies. In order to perform the clustering, seasonal patterns are distinguishable through the use of similar probability distributions. Kullback-Leibler divergence is used as a measure of the difference between the probability distributions, while the K-means algorithm is used for clustering. Finally, for the validation of the design two common methods are implemented: Nielsen Reference Model and an ARMA-GARCH model. Our comparative analysis shows that the proposed method greatly improves the precision and accuracy of the resulting wind forecasting.

1. INTRODUCTION

During the last 20 years, the electric demand has increased in a steady manner in Chile, with an actual annual rate of 6.7%. This demand is covered mainly through traditional energy sources such as fossil fuels (63%), hydroelectric (34%), and Non-Conventional Renewable Energies (3%) (NCRE) (Chilean Energy Department, 2012). This situation generates a great vulnerability to the electric supply system, which is affected by climatic factors due to global warming, or restrictions to the natural gas supplies.

In order to promote the use of clean energies, on April 1st 2008, a new law dictated the obligation for electric generation companies to supply at least 10% of the energy from NCRE by the year 2024.

For this reason, it is imperative to have tools that can predict the NCRE behavior with a certain degree of accuracy. Regarding wind generation, this need becomes more relevant since its random nature makes difficult to estimate the available energy at a determined instant.

Some efforts have been developed in order to prognosticate the behavior of wind speed, in particular, the power that can be generated in a wind park. For example the project ANEMOS (Development of A NExt generation wind resource forecasting systeM for large-scale integration of Onshore and offshore wind farmS), developed by the European Union, was created in order to study prediction models and to elaborate new methodologies (Kariniotakis, Pinson, Siebert, Giebel, & Barthelmie, 2004) (IDAE, 2007).

One characteristic that should be considered is the existence of a dominant pattern in the wind speed magnitudes in determined periods of the year, not necessarily influenced by the four seasons. If this is considered, it will allow more precise adjustments on the implemented models.

Furthermore a daily component is also present. This component is related to changes in temperature or pressure, and it manifests in similar wind speeds for determined hours of the day on relatively close days.

2. TIME SERIES PREDICTION

When dealing with wind speed data, the problem of time series prediction becomes a complex task due to its random nature. For this reason, several techniques have been developed using different approaches. On the one hand we have phenomenological approaches, where physical variables such as temperature, pressure or altitude are considered. On the other hand, we have empirical approaches. The latter type allows to learn relationships directly from data sets. Among them, statistical approaches use historic data associated with time series to characterize the frequency associated with the appearance of given characteristic patterns. This article focuses on the implementation of statistical approaches.

To deal with the data sets, some type of clustering is required. In this approach the use of tools such as the Kullback-Leibler divergence (KLD) and the K-means algorithm.

2.1. Data Clustering

System models such as the auto-regressive types, or filtering methods, can be determined with complete or partial data sets. Typically, the user is the one who determines how much data will be utilized; sometimes without solid theoretical justification. This situation brings some disadvantages. For instance, if a long time series is used, there can be parameter overestimation, and the result may not be the globally optimal. To solve this issue, it becomes necessary the use of tools that can determine differences in the patterns, as well as algorithms capable of clustering the data according to their similarities.

The combination of the KLD and the K-means algorithm allows to implement a proper clustering methodology. By using the KLD it is possible to determine the similarity between two variables, specifically by calculating the difference of their probability distributions. The use of the K-means algorithm makes possible to create data clusters based on a certain distance function.

2.1.1. Kullback-Leibler Divergence

The Kullback-Leibler divergence (Kullback, 1968) (Hershey & Olsen, 2007), also known as the entropy between two probability density functions (pdf), is commonly used in statistical analysis, as a measure of distance. The equation that defines the divergence can be expressed as follows:

$$D(f \parallel g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) \qquad (1)$$

In the particular case when the distribution functions are multivariate Gaussian functions, that we will call $\hat{f}$ and $\hat{g}$, the KLD has a closed and known shape, defined by:

$$\hat{f} \sim N\left(\mu_{\hat{f}}, \Sigma_{\hat{f}}\right) \qquad (2)$$

$$\hat{g} \sim N\left(\mu_{\hat{g}}, \Sigma_{\hat{g}}\right) \qquad (3)$$

$$D(f \parallel g) = \frac{1}{2}\begin{bmatrix} log\frac{\Sigma_{\hat{g}}}{\Sigma_{\hat{f}}} + Tr\left[\Sigma_{\hat{g}}^{-1}\Sigma_{\hat{f}}\right] - d + \cdots \\ \ldots \left(\mu_{\hat{f}} - \mu_{\hat{g}}\right)^{T}\Sigma_{\hat{g}}^{-1}\left(\mu_{\hat{f}} - \mu_{\hat{g}}\right) \end{bmatrix} \qquad (4)$$

The variables $\mu_{\hat{f}}$ and $\mu_{\hat{g}}$ correspond to the vectors of means of each distribution, $\Sigma_{\hat{f}}$ and $\Sigma_{\hat{g}}$ are the covariance matrixes and d is the vector dimension.

2.1.2. K-means Algorithm

This algorithm was proposed in (MacQueen, 1967), is one of the simplest and known clustering algorithms. It is based on a simple way to divide a given database in a predetermined *K* amount of groups. The main idea of the method is to define one centroid per each defined group. Then the data is distributed, through an iterative process that concludes when a required condition of distance is accomplished,

3. METHODOLOGY AND IMPLEMENTATION OF THE V-ARX MODEL

In this section, the implementation of a vector model for the wind speed prognosis for a 24 hour time horizon is explained. This model considers characteristics such as periodicity and seasonality.

The data used was obtained from Project EOLO, of the Geophysics Department of the University of Chile. They correspond to a series of wind speed measurements during 1990 and 1991, in a town named Punta Lengua de Vaca, Coquimbo Region, in Chile. It consists of 8760 consecutive values with a sample rate of 1 hour, and measured at an altitude of 10 meters above sea level.

The method presented in this paper has the characteristic of not using the observed data values as such. Instead, the original data series is transformed into the residual series. The use of the residual series allows the elimination of the periodical characteristics and obtains the seasonal properties required for auto-regressive models.

In order to quantify the impact of this proposal, two comparative models are implemented: Nielsen Reference model (Madsen, Pinson, Kariniotakis, Nielsen & Nielsen, (2005), and the ARMA-GARCH model (Liu, Erdem & Shi, 2011).

3.1. Daily periodicity analysis

In order to validate the existence of periodic component on the daily wind speed profiles, the autocorrelation function is calculated. This way it is possible to confirm it there exists any relation between the hours of the day and the magnitude of the wind speed at that hour.

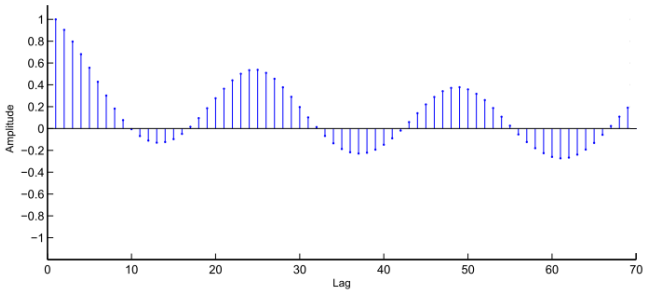

Figure 1. Auto-correlation function of the complete wind speed series data set.

Figure 1 shows that there is a strong correlation between the values of the time series. This correlation increases approximately every 24 samples; indicating the presence of a periodic component of this duration.

Knowing this, the original data is transformed into a daily vector series, starting at 1:00 am and finishing at 0:00 hours of the next day. This transformation originates 365 vectors of 24 components each. Figure 2 explains the division of the original data set while

Figure 3 shows the new distribution of data for a particular week.

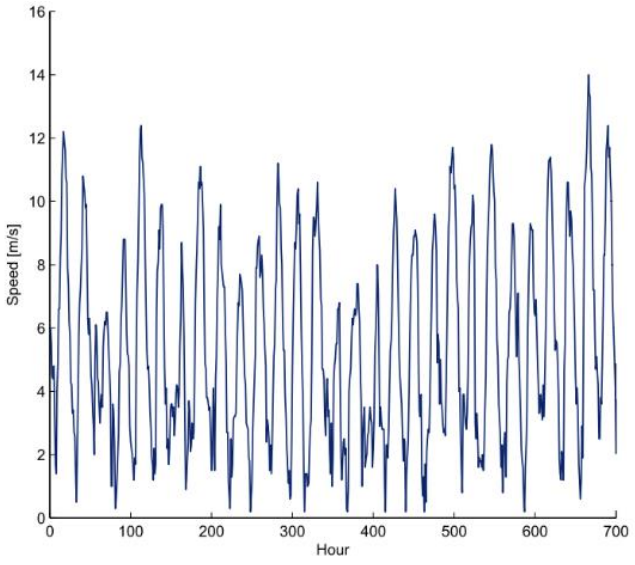

Figure 2. Original wind speed data series.

Finally each of this subseries is converted into a column vector of daily wind speeds, where each component stands for the wind speed of the specific hours for that day.

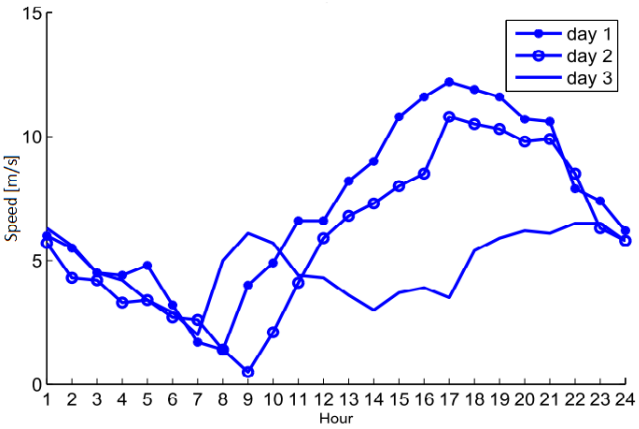

Figure 3. Example of daily wind speed values.

3.2. Seasonality analysis

In order to verify the seasonality properties of the wind speed magnitude, a clustering of the vectors, each one with similar characteristics, such as mean and variance is intended. The first step, is to arrange the daily vectors in weekly groups (7 vectors), and calculate the mean and variance of each group. This is done for all the weeks of the year. Figure 4 shows an example of the data for a particular week and how the mean and standard deviation are determined can be observed in Figure 5.

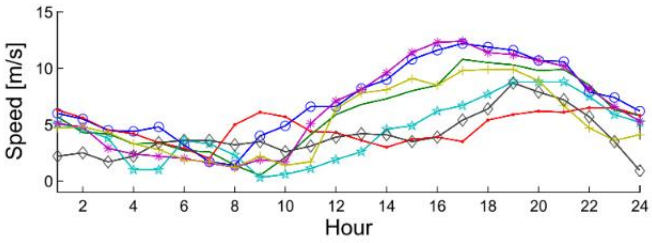

Figure 4. Example of hourly wind speed values for a complete week.

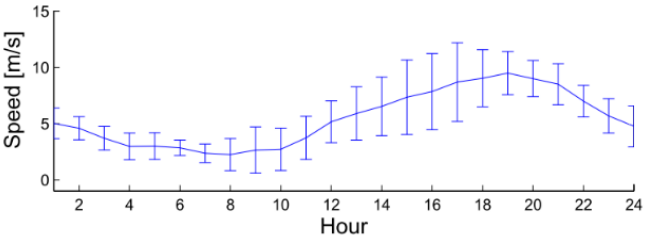

Figure 5. Wind speed mean and standard deviation.

Next step, applying the KLD to the weekly vectors of mean and variance, it is possible to determine the similarity among the different weeks of the year according to their statistical distributions. This generates a distance matrix, of dimensions $52x52$, is and each of its components indicates the difference between the two weeks. For this methodology it is assumed that the vectors are composed by independent random variables, identically distributed through Gaussian probability.

In order to cluster the data, the K-means algorithm is applied to the distance matrix obtained with the KLD. By this, all the weeks will be clustered according to the similarity of their distributions.

At this point, also an empirical component is added at the time of creating the clusters. Considering the temporal continuity needed to originate the models, some weeks are moved to different clusters, regardless of the results of K-means. In this paper, a segmentation of three clusters is obtained, and for each of them one third of the data will be used to obtain the model parameters and the remaining data is used for result validation.

3.3. Residual Series

To originate the residual data series, first it is necessary to obtain the residual vectors. They correspond to the obtained values from the transformation of each vector that contains the wind speed magnitude. These vectors are calculated by subtracting the hourly mean vector corresponding to each cluster, from the daily value vector. After this, the concatenation of these residual vectors generates the wind speed residual series. Figure 6 shows the original wind speed data and Figure 7 shows the result of this transformation.

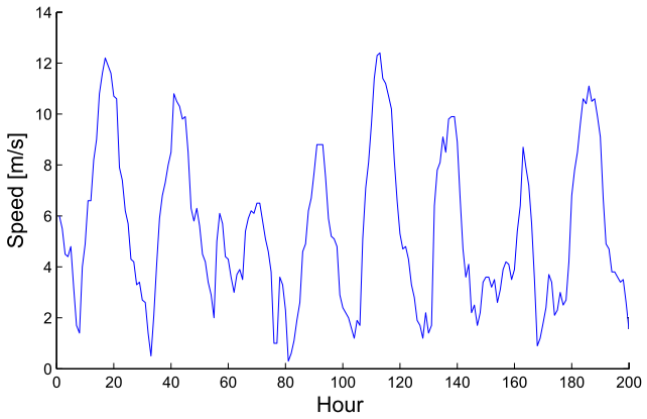

Figure 6. Original wind speed data set.

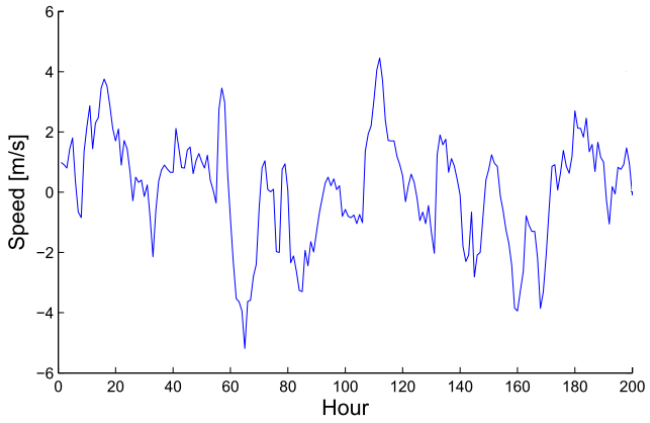

Figure 7. Wind speed residual data series.

This new time series has the characteristic of a zero mean value and also the elimination or reduction of the periodical component found in the original data set. This can be seen in Figure 8.

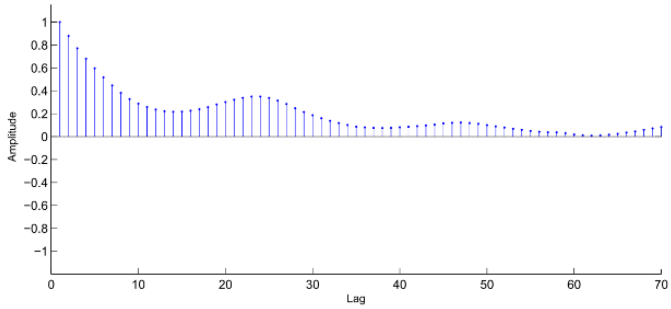

Figure 8. Auto-correlation function of residual data.

3.4. Vector Auto-Regressive Model with Exogenous Input Model

This section explains how the proposed model is obtained. It is going employ the available information until day *n*, and this will give origin to the prognosis for day *n+1*. Equation 5 defines how the prognosis is performed.

$$\bar{V}(n+1) = A_1\bar{V}_r(n) + A_2\bar{V}_r(n-1) + \cdots \\ \ldots + A_3\bar{V}_r(n-2) + \bar{V}_{ext} \quad (5)$$

The parameters are defined as follows:

- Prognosis vector $\bar{V}$: gives the result for day *n*+1. Consists of a 24 element array, where each row corresponds to one hour of the day, starting at 1 a.m. and ending at 12 a.m. of the next day.

- Residue vector $\overline{V_r}$: it has the same structure as the prognosis vector. It contains the residual values of days *n*, *n*-1, and *n*-2.
- Exogenous input $\bar{V}_{ext}$: the values of this array are the hourly mean values. It is calculated through the defined groups in the clustering step.
- Auto-regressive matrices A_i: these matrices are the corresponding auto-regressive matrices of the model, operating on the residual vectors. They are built as upper triangular matrices since a-priori it is known that the first row will originate the first prognostic result. From the auto-correlation function of the residue series, it is known that the first estimated value is the one that can capture most of the information, since it is more correlated with the nearest values. In this regard, the first row has 24 coefficients in order to use the majority of information contained on the residue vector, the second row has 23 coefficients, and so on, until the last row has only one coefficient. These coefficients must be calculated independently for each row, and using the Yule-Walker (Hershey & Olsen, 2007) equations, using the auto-covariance values of the residual series.

3.4.1. Auto-regressive parameter estimation process

1. Write the value to estimate in terms of the unknown model coefficients and the known observations that correspond.

$$\begin{aligned} V_{r,1}^{n+1} &= \alpha_{24}^{1} V_{r,1}^{n} + \alpha_{23}^{1} V_{r,2}^{n} + \alpha_{22}^{1} V_{r,3}^{n} + \\ &\dots + \alpha_{1}^{1} V_{r,24}^{n} + \alpha_{24}^{2} V_{r,1}^{n-1} + \alpha_{23}^{2} V_{r,2}^{n-1} + \alpha_{22}^{2} V_{r,3}^{n-1} \\ &+ \cdots + \alpha_{1}^{2} V_{r,24}^{n-1} + \alpha_{24}^{3} V_{r,1}^{n-2} + \alpha_{23}^{3} V_{r,2}^{n-2} \\ &+ \alpha_{22}^{3} V_{r,3}^{n-2} + \cdots + \alpha_{1}^{3} V_{r,24}^{n-2} \end{aligned} \quad (6)$$

Variable $V_{r,t}^{n+1}$: value of the wind speed prognosis for the residue series at hour *t* of day *n+1*.

Coefficient α_j^k: represents the *j-th* coefficient of matrix A_k, considered from right to left.

Variable $V_{r,t}^{n-k}(k = 0,1,2)$: value of the residual series on day *n*, *n-1*, *n-2* at hour *t*. For any other hour that requires estimation, the expression that should be used is:

$$\begin{aligned} V_{r,t}^{n+1} &= \sum_{i=t}^{24} \alpha_{25-i}^{1} V_{r,t}^{n} + \cdots \\ &\dots + \sum_{i=t}^{24} \alpha_{25-i}^{2} V_{r,t}^{n-1} + \sum_{i=t}^{24} \alpha_{25-i}^{3} V_{r,t}^{n-2} \end{aligned} \quad (7)$$

2. Both sides of the equation must me multiplied by each of the real values that are going to be used, originating a set of equations, with as many equations as parameters to be estimated.

$$\begin{aligned} V_{r,1}^{n+1} V_{r,1}^{n} &= \alpha_{24}^{1} V_{r,1}^{n} V_{r,1}^{n} + \alpha_{23}^{1} V_{r,2}^{n} V_{r,1}^{n} + \\ &\dots + \alpha_{1}^{1} V_{r,24}^{n} V_{r,1}^{n} \\ V_{r,1}^{n+1} V_{r,2}^{n-1} &= \alpha_{24}^{2} V_{r,1}^{n-1} V_{r,2}^{n} + \alpha_{23}^{2} V_{r,2}^{n-1} V_{r,2}^{n} + \\ &\dots + \alpha_{1}^{2} V_{r,24}^{n-1} V_{r,2}^{n} \\ &\vdots \\ V_{r,1}^{n+1} V_{r,24}^{n-2} &= \alpha_{24}^{3} V_{r,1}^{n-2} V_{r,24}^{n-2} + \alpha_{23}^{3} V_{r,2}^{n-2} V_{r,24}^{n-2} + \\ &\dots + \alpha_{1}^{3} V_{r,24}^{n-2} V_{r,24}^{n-2} \end{aligned} \quad (8)$$

3. Once the set of equations is determined, the expected value of each value is calculated and with them, the covariance values of the used series appear.

$$\begin{aligned} \gamma_1 &= \alpha_{24}^{1}\gamma_0 + \alpha_{23}^{1}\gamma_1 + \alpha_{22}^{1}\gamma_2 + \cdots + \alpha_{1}^{3}\gamma_0 \\ \gamma_2 &= \alpha_{24}^{1}\gamma_1 + \alpha_{23}^{1}\gamma_0 + \alpha_{22}^{1}\gamma_1 + \cdots + \alpha_{1}^{3}\gamma_{70} \\ &\vdots \\ \gamma_{72} &= \alpha_{24}^{1}\gamma_{71} + \alpha_{23}^{1}\gamma_{70} + \alpha_{22}^{1}\gamma_{69} + \cdots + \alpha_{1}^{3}\gamma_0 \end{aligned} \quad (9)$$

4. Finally, the coefficient values are determined.

$$\begin{bmatrix} \alpha_{24}^{1} \\ \alpha_{23}^{1} \\ \alpha_{22}^{1} \\ \vdots \\ \alpha_{1}^{3} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{71} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{70} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{69} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{71} & \gamma_{70} & \gamma_{69} & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \vdots \\ \gamma_{72} \end{bmatrix} \quad (10)$$

With this procedure, all the rows are calculated taking into consideration that each time the estimation moves away an hour, one auto-regressive coefficient less is going to be used.

3.5. Methodology step-by-step summary

The proposed methodology, can be summarized following the next steps:

1. From raw data, create the conventional wind speed time series.
2. Divide the conventional time series into weekly groups. Each of these groups should consist on the hourly wind speed value of each day.
3. Calculate the mean and variance for each of the groups created on step 2.
4. Apply the KLD to the vector data created on Step 3, this will create a distance matrix.
5. Using K-means algorithm, create clusters using the KLD results.
6. Create the V-ARX model.
7. Generate prognosis results for the V-ARX model and the other techniques.
8. Measure the performance indicators.

9. Compare the proposed V-ARX model with the reference techniques: Nielsen Reference model and ARMA-GARCH.

4. RESULTS

In this section, we present the obtained results with the proposed V-ARX model as well as the Nielsen Reference model and the ARMA-GARCH model that were intended for comparison purposes.

4.1. Seasonal Groups Determination

The intention is to distribute the weeks of the year into different groups that present similar wind speed daily patterns. As mentioned previously, this is done by using the KLD to obtain the matrix of distances. Then, the K-means algorithm is used to create the data clusters, independently of their temporal location during the year. For this reason, an empirical criteria is employed in order that these groups of data are formed by temporarily correlated data. Figure 9 shows the results of the clustering process.

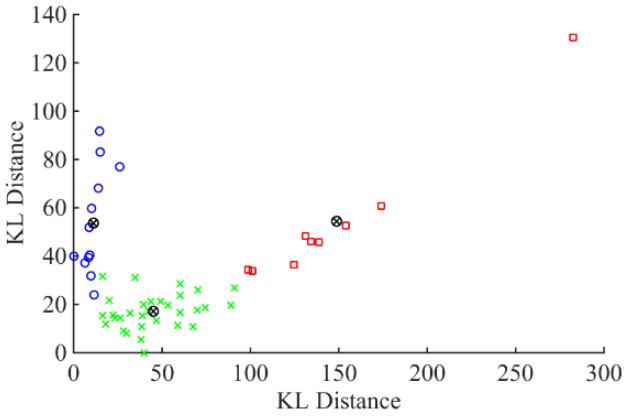

Figure 9. K-means clustering results.

In this case, three clusters were used to divide the available data, since a larger amount of groups will originate clusters with few weeks, opposing to the intention of the research, which is to minimize the amount of necessary models. The green crosses represent the data assigned to Cluster A, the blue circles to Cluster B and the red squares to Cluster C. All the weeks can be arranged as shown in Table 1.

Table 1. Cluster Distribution of the Weeks using K-means

Week	1	2	3	4	5	6	7	8	9	10	11	12	13
Cluster	A	A	A	A	A	A	B	A	B	B	C	B	B
Week	14	15	16	17	18	19	20	21	22	23	24	25	26
Cluster	C	B	B	C	B	C	C	C	B	B	C	B	C
Week	27	28	29	30	31	32	33	34	35	36	37	38	39
Cluster	C	B	B	C	B	B	B	B	C	B	B	B	B
Week	40	41	42	43	44	45	46	47	48	49	50	51	52
Cluster	B	B	A	A	B	A	B	A	B	B	A	B	B

The results obtained with K-means are used as the base structure to create new groups and with the support of an empirical analysis they are defined as:

- Group 1: weeks 1 to 13.
- Group 2: weeks 14 to 30.
- Group 3: weeks 31 to 52.

The characteristic pattern of mean and deviation for the three groups are shown on figures 10 through 12.

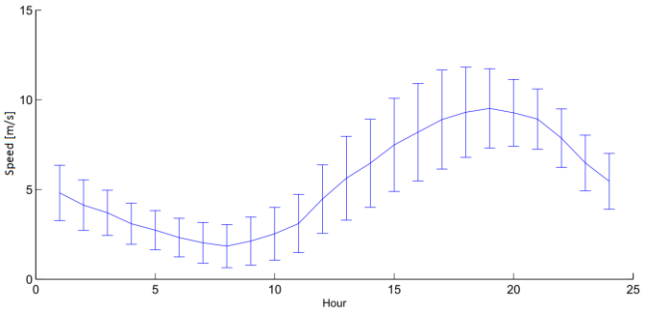

Figure 10. Mean and standard deviation value for Group 1.

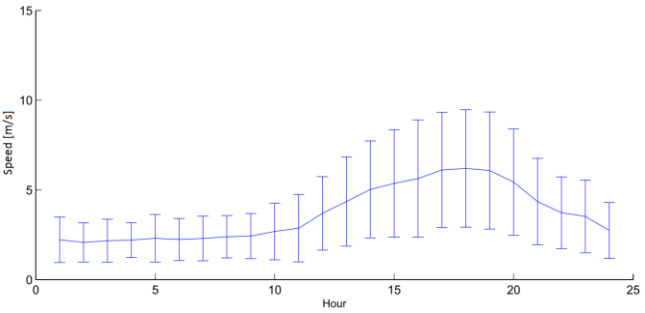

Figure 11. Mean and standard deviation value for Group 2.

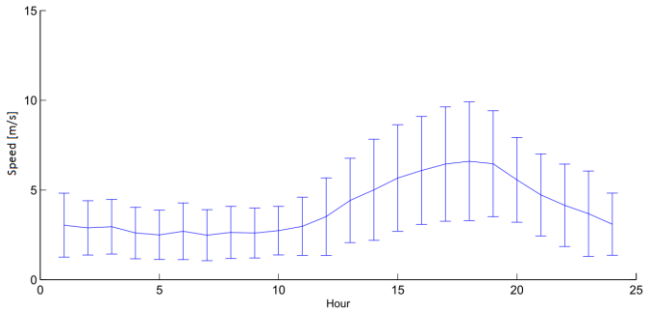

Figure 12. Mean and standard deviation value for Group 3.

The next figures show the autocorrelation function of the original and the residual series respectively, demonstrating that the foundation of the model are still present: a highly marked periodicity of the original series and in the residual series a small periodic component with low correlation values.

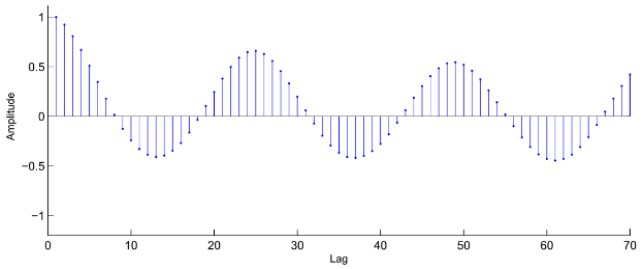

Figure 13. Autocorrelation function of the original data set of Group 1.

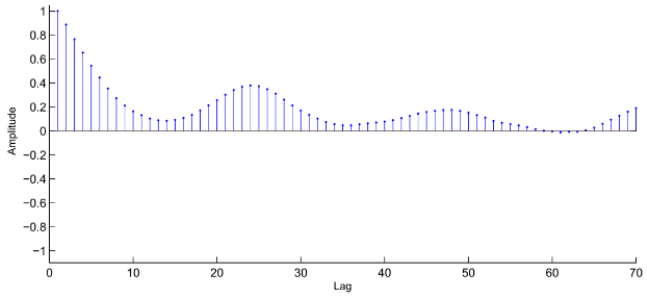

Figure 14. Autocorrelation function of the residual data set of Group 1

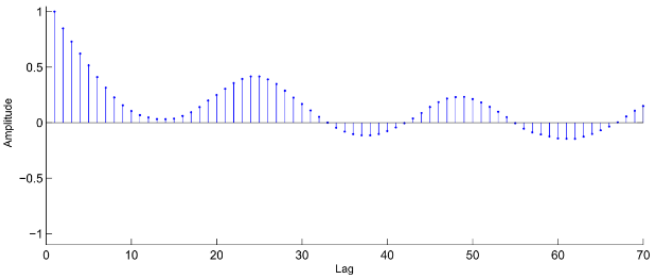

Figure 15. Autocorrelation function of the original data set of Group 2.

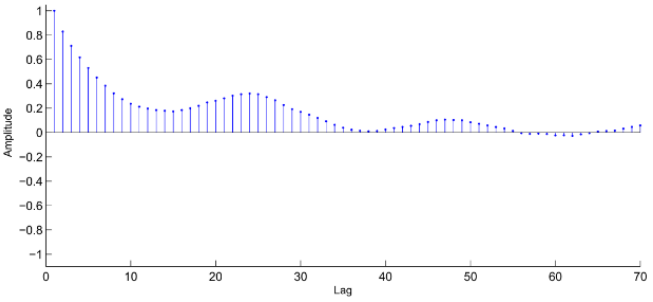

Figure 16.Autocorrelation function of the residual data set of Group 2.

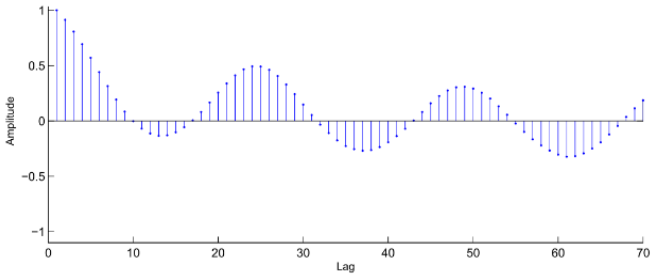

Figure 17.Autocorrelation function of the original data set of Group 3.

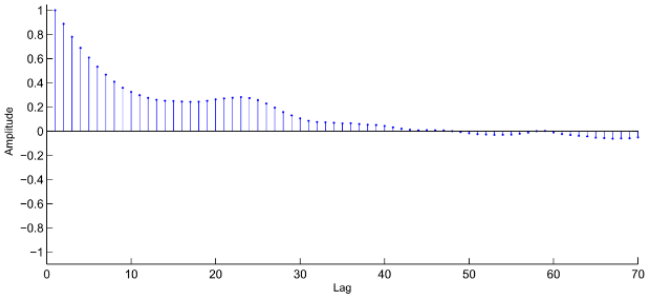

Figure 18. Autocorrelation function of the residual data set of Group 3.

4.2. Prognosis Results

Using the previously explained methodology, the V-ARX model is implemented. The parameters are estimated using one third of the available data, and the rest is used for validation purposes. Figures 19-21 show the actual data and the prognosis results for the first 7 days (168 hours), for the validation data set. It is important to keep in mind that the prognosis time horizon is 24 hours, and after this period of time elapses, the values that enter the model for the next prognosis correspond to the real values.

The results show that the V-ARX model is capable of following the original series trend, mainly because the use of the mean speed pattern, represented in the exogenous input. However, due to the strong influence of this parameter on the estimated values, it is possible that if the real value varies too much with respect to the mean values, the prognosis error will increase since the auto-regressive part will not be able to estimate correctly such magnitude variation. This can be seen between hours 1 to 40 of Group 2 and 70 to 110 of Group 3, where the real values are smaller than the mean values of the data series.

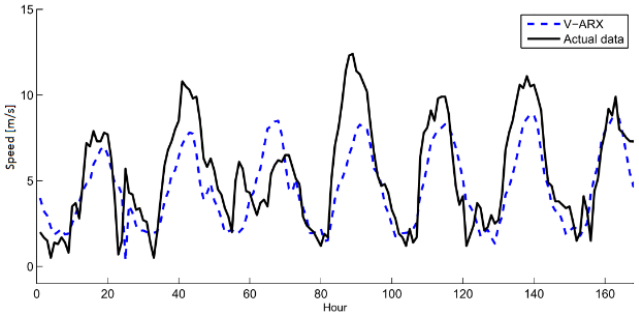

Figure 19. V-ARX model wind speed estimation of Group 1.

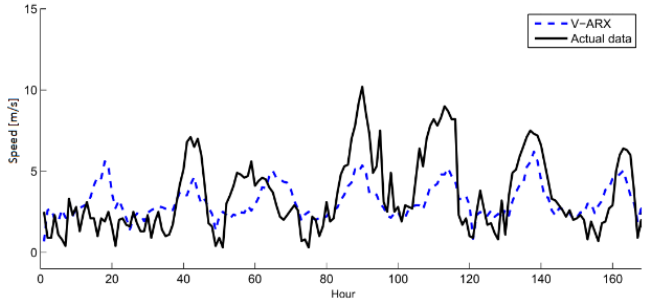

Figure 20. V-ARX model wind speed estimation of Group 2.

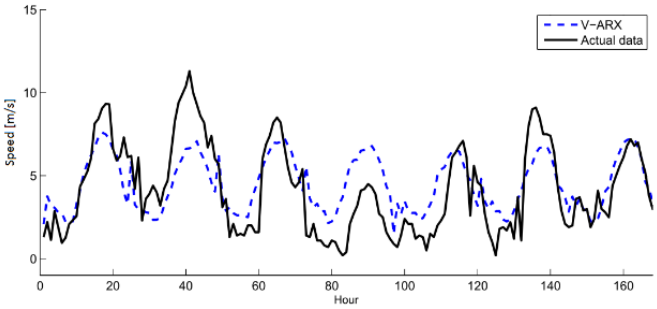

Figure 21. V-ARX model wind speed estimation of Group 3.

For the Nielsen Reference model, a 24 hour prognosis time horizon is also employed. The training and validation data are also the same as before. Figures 22-24 show the results for the first 7 days of the validation data.

The results are clear to show that the Nielsen Reference model does not provide a good performance. The reason is that the model equation requires a mean component, and in this case it uses the historic mean instead of the hourly mean. In this regard, the predictions made for a relatively far horizon will tend to emulate the mean value since it becomes more difficult to capture the changes of the auto-correlation values of the series. Contrary, if looking closely the first 3 hours of each day, the results of the prognosis is really close to the actual data. This is caused by the strong correlation between hours that are near to the prior time instant when the real value is used to estimate.

Finally, an ARMA(p, q)-GARCH(1,1) model is implemented in a similar manner as the previous models. In this case, the ARMA-GARCH model is applied over the residual series, so an external variable is present. It corresponds to the mean value of the hour at which the prognosis is being done. The parameters p and q used for each group is different and they are obtained by calculating the minimum mean square error when varying p and q between 1 and 6. Figures 25-27 show the results for the first seven days of the validation data set.

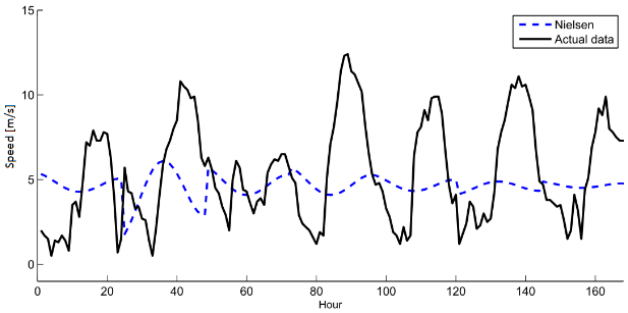

Figure 22. Nielsen Reference model wind speed estimation of Group 1.

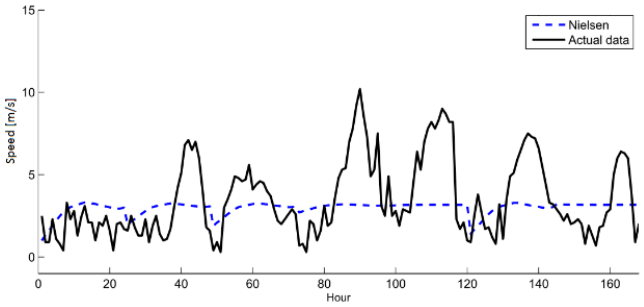

Figure 23. Nielsen Reference model wind speed estimation of Group 2.

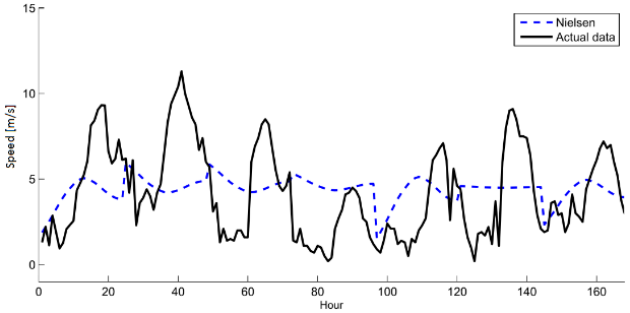

Figure 24. Nielsen Reference model wind speed estimation of Group 3.

Since the ARMA-GARCH model is designed with the residual series and using the hourly mean values, it is able to estimate the trend of the series satisfactorily. Once again the component of mean values of the wind speed plays a key role when doing the prognosis. The results are very similar to the V-ARX model. The comparison of the models will allow to stablish which model is more accurate.

In case the ARMA-GARCH model is developed using the observed data and not the residual series, the results will be very similar to the Nielsen Reference model case.

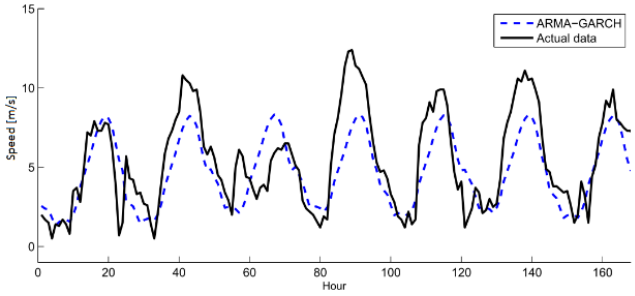

Figure 25. ARMA-GARCH model wind speed estimation of Group 1.

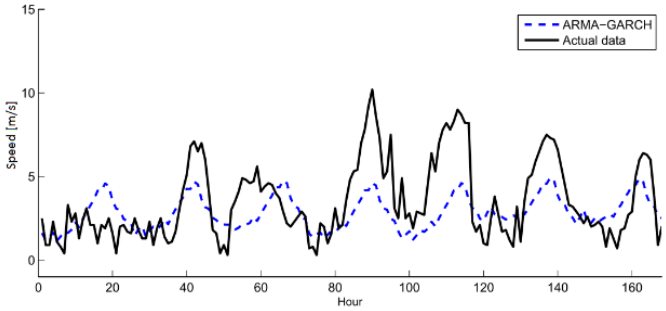

Figure 26. ARMA-GARCH model wind speed estimation of Group 2.

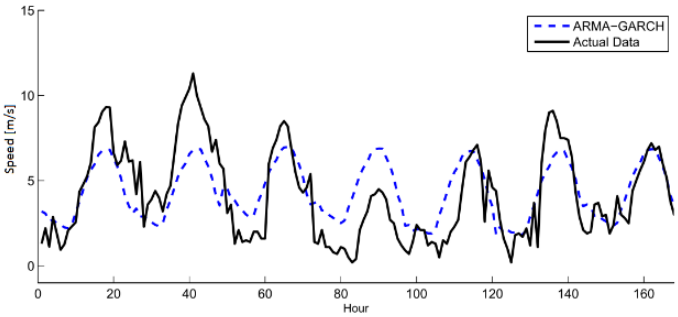

Figure 27. ARMA-GARCH model wind speed estimation of Group 3.

4.3. Performance Indicators

In this approach the prediction error of the proposed method will be compared to the other techniques using the following indicators: mean square error (MSE), mean relative percentage error (MRPE), mean relative effective percentage error (MREPE) and the mean percentage energy error (MPEE). The prediction error can be defined as the difference between the real value of the series at certain time instant and the estimated value by the model at that time instant. So the prediction error for a given future time instant can be calculated as:

$$e(t+k|t) = V(t+k|t) - \hat{V}(t+k|t) \qquad (11)$$

The MSE is the result of calculating the average of the squares of the prediction errors. In this case, the bigger errors are given more weight at the time they are averaged, since the values are squared. Next, the MRPE is the result of calculating the prediction error as a percentage of the real value. Then, the MREPE can be calculated as the percentage of the prediction error with respect to the average of the real value. Finally, the MPEE is obtained by averaging the energy between the prediction error and the energy of the real value. It is possible to define the previously mentioned performance metrics, using Eq. (11), as follows:

$$MSE(k) = \frac{1}{k}\sum_{i=1}^{k} e(t+k|t)^2 \quad (12)$$

$$MRPE(k) = \frac{1}{k}\sum_{i=1}^{k} \frac{|e(t+k|t)|}{|V(t+k|t)|} * 100 \quad (13)$$

$$MREPE(k) = \frac{1}{k}\sum_{i=1}^{k} \frac{|e(t+k|t)|}{\left|\frac{1}{k}\sum_{i=1}^{k} V(t+k|t)\right|} * 100 \quad (14)$$

$$MPEE(k) = \frac{1}{k}\sum_{i=1}^{k} \frac{e(t+k|t)^2}{V(t+k|t)^2} * 100 \quad (15)$$

The indicators are calculated after a 24 hour time horizon prognosis. The following tables show the obtained results of the different prognosis errors.

Table 2. Prognosis errors using the proposed V-ARX model

Group	MSE	MRPE	MREPE	MPEE
1	3.86	68.1	42.9	13.25
2	4.79	89.83	61.19	25.53
3	5.23	83.94	50.49	18.15

Table 3. Prognosis errors using the Nielsen Reference model

Group	MSE	MRPE	MREPE	MPEE
1	8.74	118.31	64.55	30
2	6.08	96.31	68.95	32.42
3	8.1	102.31	62.82	28.1

Table 4. Prognosis errors using the ARMA-GARCH model.

Group	MSE	MRPE	MREPE	MPEE
1	4.22	73.91	44.85	14.48
2	5.24	91.61	63.95	27.89
3	5.6	87.82	52.24	19.43

4.4. Model Comparison

In order to compare the obtained results with each model and determine if there is an improvement on the prognosis, the following indicator is implemented. It is called IMP (from improvement) and it indicates the percentage of improvement of the obtained prognosis when compared to a reference model.

$$IMP = \frac{e(k)_{ref} - e(k)}{e(k)_{ref}} * 100 \quad (16)$$

In this case, $e(k)_{ref}$ represents the error obtained with the reference model for a k-step prediction horizon, and $e(k)$ denotes the error of the model under study. In simple words if the value of IMP is positive the studied model has a better performance that the reference model.

The following tables show the results of calculating the improvement rate when comparing the V-ARX model with the Nielsen Reference model and then with the ARMA-GARCH model.

From the IMP results it is clear that the V-ARX model has a better performance when compared to the other models, regardless of the type of error that is being considered.

Table 5.Improvement percentage index of comparing V-ARX model with the Nielsen Reference Model.

Group	MSE	MRPE	MREPE	MPEE
1	55.84	42.44	33.55	55.84
2	21.25	6.72	11.26	21.25
3	35.42	17.99	19.67	35.48

Table 6.Improvement percentage index of comparing V-ARX model with the ARMA-GARCH Model.

Group	MSE	MRPE	MREPE	MPEE
1	8.57	7.87	4.36	8.53
2	8.46	1.94	4.36	8.46
3	6.6	4.42	3.35	6.6

The proposed V-ARX model is able to confront the variability of the hourly mean values of the wind speed, and for this reason it is more precise when doing the prognosis for Group 1 (which has a more variable pattern). Furthermore, when comparing the V-ARX model with the Nielsen Reference model there is a notorious improvement on the precision of the prognosis. This improvement is mainly due to the use of the hourly pattern present in the wind speed series making possible to use relevant information for faraway horizons where the uncertainty is larger. This issue cannot be accounted in an effective manner with the Nielsen, causing a poor prognosis performance for the 24 step horizon.

Likewise, when comparing the V-ARX model with the ARMA-GARCH there is a slight improvement. This

improvement cannot be considered as a result of just using the mean speed pattern since both models use it. Therefore the difference is the result of how the auto-regressive part is solved, resulting that the proposed matrix structure has a better performance that the linear manner. One of the reasons of this improvement is that the V-ARX has all the information presented at once to obtain the prognosis vector without incurring in errors associated to the reincorporation of estimated values to the model, while the ARMA-GARCH model uses the prognosis values as feedback, increasing the existing level of uncertainty and the estimation error.

5. CONCLUSIONS

The presented research consists on the development of a dynamic vector model applied to wind speed prognosis for a 24 hour time horizon. It was designed in such a way that is capable of capturing the periodic characteristics of the wind, daily and seasonal. This demonstrates that considering all the information when building a model the results are improved.

Another fact to highlight is that with just one iteration a 24 hour prognosis horizon is performed by using all the relevant information at once. If a linear model is employed to perform this type of prognosis, the calculation has to be done 24 times. In this case, the results have to go through a feedback process into the model equation, increasing the associated errors with each iteration. In this context, it is important to recognize the utility of using the residual series and the mean values of the speed patterns, allowing to perform a longer horizon prognosis and obtaining estimation results closer to the real values.

The values obtained for the *IMP* indicator demonstrate that the presented vector model gives better results that the two models used for comparison. It is possible to conclude that the proposed design, has a great potential since it only uses an AR process to solve the auto-regressive part and is able to overcome two well-known processes. In this regard, if the vector technique is adapted to other algorithms that are more precise, the overall results can be improved, resulting into more accurate estimations.

The way the groups are defined is not present on other researches, and the obtained results are promising. The fact of having used K-means clustering and Kullback-Leibler divergence allowed to perform a good and mathematically founded, identification of the different wind patterns throughout the year, making possible to create a more precise model for each of the groups. Having different models depending on the amount of groups allowed to have better results rather than having worked with just one model for the whole year.

ACKNOWLEDGEMENTS

This work has been partially supported by FONDECYT Chile Grant Nr. 1140774, and the Advanced Center for Electrical and Electronic Engineering, AC3E, Basal Project FB0008, CONICYT. The work of Jorge Silva was supported by CONICYT-Chile, FONDECYT Chile Grant Nr. 1140840. The work of Aramis Pérez was supported by the University of Costa Rica (Grant for Doctoral Studies) and CONICYT-PCHA/Doctorado Nacional/2015-21150121.

REFERENCES

Chilean Energy Department (2012), *National Energy Strategy 2012-2030*, Santiago, Chile.

Kariniotakis, G., Pinson, P., Siebert, N., Giebel, G., & Barthelmie, R. (2004). The state of the art in short-term prediction of wind power-from an offshore perspective. *In Proceedings of 2004 SeaTechWeek*. October 20-21, Brest, France.

Instituto para la Diversificacion y Ahorro de la Energia (IDAE) (2007). ANEMOS. Estudio sobre la Prediccion Eolica en la Union Europea. Madrid, Spain.

Kullback, S. (1968). *Information theory and statistics*. Mineola, New York. Dover Publications.

Hershey, J. R., & Olsen, P. A. (2007). Approximating the Kullback Leibler divergence between Gaussian mixture models. *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, pp. IV-317-IV-320. Honolulu, Unites States of America. doi: 10.1109/ICASSP.2007.366913.

MacQueen, J., (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Statistics, 281-297, University of California Press, Berkeley, CA, USA. http://projecteuclid.org/euclid.bsmsp/1200512992.

Madsen, H., Pinson, P., Kariniotakis, G., Nielsen, H. A., & Nielsen, T. (2005). Standardizing the performance evaluation of short term wind power prediction models. *Wind Engineering*, vol. 29, no. 6, pp. 475-489.

Liu, H., Erdem, E., & Shi, J. (2011). Comprehensive evaluation of ARMA–GARCH (-M) approaches for modeling the mean and volatility of wind speed. *Applied Energy*, 2011, vol. 88, no 3, pp. 724-732.

BIOGRAPHIES

M. Sc. Aramis Pérez is a Research Assistant at the Lithium Innovation Center (Santiago, Chile) and Professor at the School of Electrical Engineering at the University of Costa Rica. He received his B.Sc. degree (2002) and Licentiate degree (2005) in Electrical Engineering from the University of Costa Rica. He received his M.Sc. degree in Business Administration with a General Management Major (2008) and also he is a M.Sc. candidate in Industrial Engineering from the same university. Currently he is a doctorate student at the Department of Electrical Engineering at the University of Chile under Dr. Marcos E. Orchard supervision. His research interests include parametric/non-parametric modeling, system identification, data analysis, machine learning and manufacturing processes.

B. Sc. Francisco Cornejo received the B.Sc. degree in Electrical Engineering from Universidad de Chile, Santiago, Chile, in 2012.

Dr. Marcos E. Orchard is Associate Professor with the Department of Electrical Engineering at Universidad de Chile and was part of the Intelligent Control Systems Laboratory at The Georgia Institute of Technology. His current research interest is the design, implementation and testing of real-time frameworks for fault diagnosis and failure prognosis, with applications to battery management systems, mining industry, and finance. His fields of expertise include statistical process monitoring, parametric/non-parametric modeling, and system identification. His research work at the Georgia Institute of Technology was the foundation of novel real-time fault diagnosis and failure prognosis approaches based on particle filtering algorithms. He received his Ph.D. and M.S. degrees from The Georgia Institute of Technology, Atlanta, GA, in 2005 and 2007, respectively. He received his B.S. degree (1999) and a Civil Industrial Engineering degree with Electrical Major (2001) from Catholic University of Chile. Dr. Orchard has published more than 50 papers in his areas of expertise.

Dr. Jorge F. Silva is Associate Professor at the Department of Electrical Engineering, University of Chile, Santiago, Chile. He received the Master of Science (2005) and Ph.D. (2008) in Electrical Engineering from the University of Southern California (USC). He is IEEE member of the Signal Processing and Information Theory Societies and he has participated as a reviewer in various IEEE journals on Signal Processing. Jorge F. Silva is recipient of the Outstanding Thesis Award 2009 for Theoretical Research of the Viterbi School of Engineering, the Viterbi Doctoral Fellowship 2007-2008 and Simon Ramo Scholarship 2007-2008 at USC. His research interests include: non-parametric learning; sparse signal representations, statistical learning; universal source coding; sequential decision and estimation; distributive learning and sensor networks.