Applying data normalization for the Solar Radiation Modelling

JOSÉ G. HERNÁNDEZ-TRAVIESO[1], CARLOS M. TRAVIESO[1], JESÚS B. ALONSO[1], MALAY KISHORE DUTTA[2].
[1]Signal and Communications Department, Institute for Technological Development and Innovation on Communications (IDeTIC).
University of Las Palmas de Gran Canaria.
Campus Universitario de Tafira, sn, Ed. de Electrónica y Telecomunicación, Pabellón B, Desp. 111, E35017, Las Palmas de Gran Canaria.
SPAIN.
[2]Department of Electronics and Communications Engineering,
Amity University, Noida.
INDIA.
jose.hernandez149@alu.ulpgc.es,{carlos.travieso,jesus.alonso}@ulpgc.es, mkdutta@amity.edu

Abstract: To normalize data gives the opportunity to reduce samples dispersion in order to obtain a better result in several investigations. This work tries to verify this hypothesis using meteorological data and different normalization methods. Meteorological data were provided by State Meteorological Agency (AEMET) depending of Spanish Government from meteorological stations located in Gran Canaria and Tenerife (Canary Islands, Spain), working with data of solar radiation and applying normalization to solar radiation prediction.

Key-Words: Data normalization; climate application; weather prediction, solar radiation prediction; reduce samples dispersion.

1 Introduction

If we have a dataset with various values with large differences between them and we want to reduce those differences to thereby obtain a more compact dataset that allows us to work with less stress into the system, normalization could be an option.

Normalization pretends to reduce dispersion of data collected in order to obtain better results in the experiments realized with them.

At this point, we introduce only a few works related to the topic.

In 2011, Huang, Zhou, Zhang, Zhang and Li, used normalized meteorological data to predict Malaria in Central China [1]. Vašak, Gulin, Čeović, Nicolić, Pavlović and Perić used normalized data to obtain a prediction of electrical power delivery of a photovoltaic panel, in 2011 [2].

In 2013, Zeng and Quiao used normalized data of meteorological variables to obtain a solar power prediction [4]. Ma. Li, Yang, Du and Wang, used in 2014 normalized meteorological data to make a prediction model for short-term wind farm output power. In 2014, Hernández-Travieso, Travieso, Alonso and Dutta [6], using ANN to modelling solar radiation obtaining a MAE of 0.04 kWh/m^2 for the estimation of solar energy generation.

According the state-of-the-art, this work will contributes to spread this line, using three different normalization blocks in order to obtain which one offers better results.

By means of applying normalization block for the solar radiation modelling this contribution could be possible. For that reason, this study used data collected by meteorological stations located on the island of Gran Canaria (GC) at the Gran Canaria Airport (Spain) and on the island of Tenerife (TF) at Tenerife Sur Airport (Spain) controlled by Spanish Meteorological Agency (*Agencia Estatal de Meteorología*, AEMET- ESPAÑA).

2 Normalization block

In this study, three different types of normalization have been tested:

- Normalization between maximum and minimum values [-1 +1].
- Null mean normalization.
- Normalization by decades.

Normalized results have been compared to results obtained with un-normalized data test to check the goodness of each method, allowing us to determine which one gives better results.

Each one of these types will be explained below.

2.1 Normalization between maximum and minimum values [-1 +1]

By mean of this normalization method, all input data will have a value between -1 and +1. This normalization reduces significantly the dispersion of samples if we have values with large variations values.

In this case, the equation (1) is applied for the normalization, and gives also the opportunity to apply to several data test the same normalization pattern applied to data training, not only at normalization but also in the opposite process;

$$\left(2 * \frac{\max\_value - sample_i}{\max\_value - \min\_value}\right) - 1 \qquad (1)$$

Where:

- *max_value* is the maximum value of un-normalized input data.
- *min_value* is the minimum value of un-normalized input data.
- $sample_i$ is the sample to normalize.

This function applies normalization by rows to the input data matrix, so each row will have a maximum and a minimum value and the rest of the values will be normalized taking into account this maximum and minimum values.

2.2 Null mean normalization

In this case, equation (2) was used to normalize:

$$\left(\frac{i_{SN}}{\overline{\chi}\,(i_{SN})}\right) - 1 = i_N \qquad (2)$$

Where:

- i_{SN} stand for data input un-normalized.
- $\overline{\chi}\,(i_{SN})$ stand for mean value of un-normalized input data.
- i_N stand for data input normalized.

To return to de original data once the experiment has been done it is necessary to do the opposite process using equation (3):

$$(e_N + 1) \cdot \overline{\chi}\,(i_{SN}) = e_{SN} \qquad (3)$$

Where:

- e_N stand for data output normalized.
- e_{SN} stand for data output un-normalized.

Data inputs acquire values into the range of -1 to +1, but the difference to the previous method explained in 2.1, in this case there is no common pattern to normalize and to de-normalize data.

2.3 Normalization by decades

To perform this normalization, data inputs were normalized using different decade values in order to reduce dispersion of samples following equation (4) to normalize and equation (5) to de-normalize.

$$\frac{i_{SN}}{d} = i_N \qquad (4)$$

$$e_N \cdot d = e_{SN} \qquad (5)$$

Where:

- d stand for the decade value used.

3 Experimental methodology and results

3.1 Data filtering

AEMET provides a payment database including data relative to solar radiation, wind speed, temperature, meteor and humidity and due to the loyalty of the institution. Besides, it gives the opportunity to do the study in two different geographical locations

Files has .xls format, with one file per phenomenon with both stations together (GC and TF), for a period of time of five years, from 2003 to 2007.

Data was collected hourly for the phenomena involved in this work (precipitation, temperature, wind speed and solar radiation), other phenomena were rejected according to criteria of data absence due to maintenance or failure of the sensor, no data collection hourly during 24h or importance of the phenomenon in order to obtain a result.

Then it is necessary to adjust the database to the environment used in this work (Matlab). For the prediction system, Artificial Neural Network (ANN) was used. An ANN is a machine that is designed to model the way in which the brain performs a particular task or function of interest [7].

The dimension of each file is shown on Table 1.

METEOROLOGICAL PHENOMENA	.xls DIMENSION AEMET (rows x columns)
Humidity	3653 x 37
Meteor	121 x 52
Cloudiness	3653 x 18
Precipitation	3653 x 35
Radiation	1900 x 26
Temperature	3653 x 37
Wind speed	3652 x 60

Table 1: Files dimensions.

ISBN: 978-1-61804-280-4

3.2 Experiments

Once the data was adjusted to the correct way to introduce into ANN in Matlab, the number of samples of the resulting file is shown on Table 2 and then is introduced into Matlab.

YEAR	SAMPLES PER STATION	
	GC	TF
2003	-----	1344
2004	4064	2960
2005	1296	3104
2006	1952	5424
2007	3232	5568

Table 2: Samples per meteorological station.

Heuristic methods were used to prove normalization using wider reference values with data from AEMET —precipitation in tenths of a millimeter, temperature in tenths of °C, wind speed in kilometers per hour and solar radiation in tenths of kilojoules per square meter — or more precise ones —precipitation in millimeters, temperature in °C, wind speed in meters per second and solar radiation in kilowatts hour per square meter —.

Based on the previous data, the sequential line of experiments are:

1) *Using data from AEMET.*
2) *Using more precise data.*
3) *Using different phenomena linked: data fusion.*

All normalization method has been applied to the experiments previously explained. In all of them, the values used to obtain predictions are past values of the phenomena involved in the experiment.

The system was previously adjusted to determine the right configuration of the ANN and it was trained using Neural Network Toolbox of Matlab. The year 2006 was used to train the ANN due to the number of samples included on it, which enables to train with low and high number of samples depending on the island. The test was done with the remaining years.

According to the results, and always based on heuristic methods, it will be verified the best option to normalize data in order to obtain better precision on solar radiation prediction.

4 Results

Results are presented in tables with the statistical parameters of minimum, maximum, mean absolute error, standard deviation and mean square error for both stations (GC and TF), between the true value measured by the stations and the predicted value obtained with the ANN. The statistical that gives the goodness of prediction is the MAE. In addition, the times of training and test per sample are given in seconds (sec).

According to the experiments, best results were reached using data fusion, not only for data from AEMET, but also for more precise units. For that reason, these are the results presented below. In addition, results obtained applying normalization blocks 2.1 and 2.2 had le to such bad results that it is best not explain thoroughly at this point, only comment that the best result using these blocks differed quite of the result obtained using block 2.3. An example of this was the result obtained using 2.1 that reached a MAE of 45 kWh/m^2 a result absolutely unacceptable.

4.1 GC station

At GC station, using data fusion of radiation and hour from AEMET, results were as follows on Table 3 and Fig. 1. In this experiment, best results were reached using normalization by decade 10.

STATISTICAL PARAMETERS		PREDICTION ERROR IN KILOWATTS HOUR PER SQUARE METER		
		YEAR 2004	YEAR 2005	YEAR 2007
Minimum		5.55×10^{-6}	4.80×10^{-5}	1.58×10^{-5}
Maximum		0.48	0.40	0.65
MAE		0.10	0.09	0.10
Standard deviation		0.83	0.82	0.08
Mean square error		0.13	0.13	0.13
TIME (sec)	TRAIN	2.65 x 10		
	TEST per sample	4.75×10^{-3}		

Table 3: Results using radiation and hour normalized from AEMET GC.

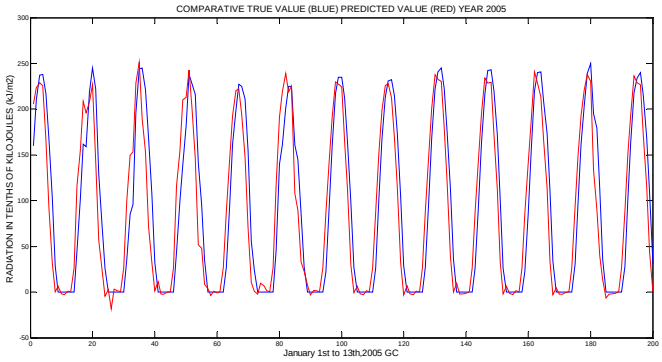

Fig. 1: Comparative with radiation and hour AEMET GC using data normalized by 10.

In 2005 with an error of 0.09 kWh/m^2, best result was reached.

Using more precise units with data fusion of radiation and hour, and normalizing by decade 10,

best results obtained was an error of 0.09 kWh/m^2 for the year 2005. As seen on Table 4 and Fig. 2:

STATISTICAL PARAMETERS		PREDICTION ERROR IN KILOWATTS HOUR PER SQUARE METER		
		YEAR 2004	YEAR 2005	YEAR 2007
Minimum		1.48×10^{-5}	1.10×10^{-5}	2.87×10^{-6}
Maximum		1.96	0.65	1.37
MAE		0.10	0.09	0.10
Standard deviation		0.10	0.09	0.10
Mean square error		0.14	0.13	0.14
TIME (sec)	TRAIN	2.66 x 10		
	TEST per sample	4.68×10^{-3}		

Table 4: Results using radiation and hour normalized GC using more precise units.

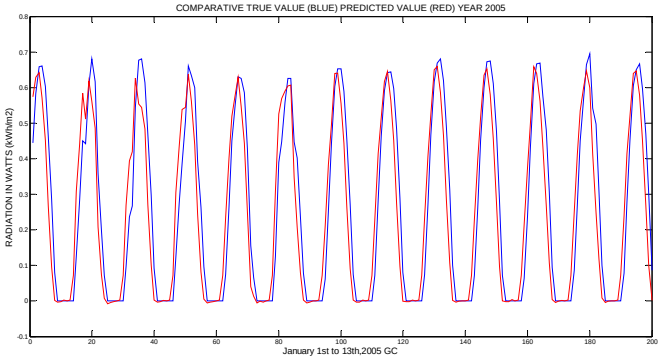

Fig. 2: Comparative with radiation and hour GC using data normalized by 10.

Data fusion of radiation and hour reaches better results than any other combination of phenomena in the study at GC station.

4.2 TF station

Using data from AEMET and data fusion of radiation and hour, best results at TF station are shown at Table 5 and Fig. 3. In this case, normalization by decade 10 was used.

STATISTICAL PARAMETERS		PREDICTION ERROR IN KILOWATTS HOUR PER SQUARE METER			
		YEAR 2003	YEAR 2004	YEAR 2005	YEAR 2007
Minimum		1.07×10^{-4}	6.67×10^{-6}	5.55×10^{-5}	5.55×10^{-5}
Maximum		0.28	0.36	0.31	0.31
MAE		0.06	0.09	0.09	0.09
Standard deviation		0.06	0.07	0.07	0.07
Mean square error		0.09	0.11	0.11	0.09
TIME (sec)	TRAIN	3.86 x 10			
	TEST per sample	4.01×10^{-3}			

Table 5: Results using radiation and hour normalized from AEMET TF.

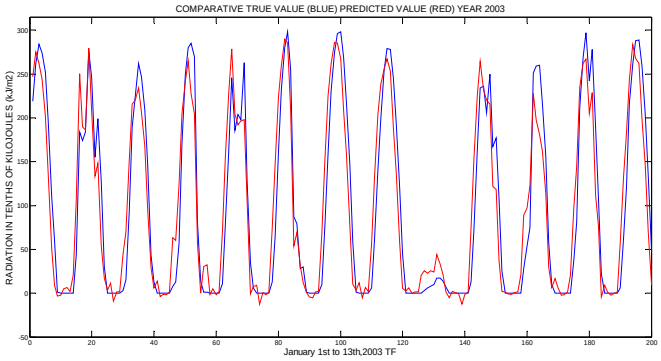

Fig. 3: Comparative with radiation and hour AEMET TF using data normalized by 10.

In 2003 with an error of 0.06 kWh/m^2, best result was reached using this configuration in the experiment.

Using more precise units, data fusion of radiation and hour and normalization by decade 20, an error of 0.06 kWh/m^2 was the best result. As seen on Table 6 and Fig. 4:

STATISTICAL PARAMETERS		PREDICTION ERROR IN KILOWATTS HOUR PER SQUARE METER			
		YEAR 2003	YEAR 2004	YEAR 2005	YEAR 2007
Minimum		2.53×10^{-6}	1.25×10^{-5}	6.87×10^{-6}	6.87×10^{-6}
Maximum		0.27	0.32	0.33	0.33
MAE		0.06	0.08	0.08	0.08
Standard deviation		0.07	0.07	0.07	0.07
Mean square error		0.09	0.11	0.11	0.08
TIME (sec)	TRAIN	3.84 x 10			
	TEST per sample	3.99×10^{-3}			

Table 6: Results using radiation and hour normalized TF using more precise units.

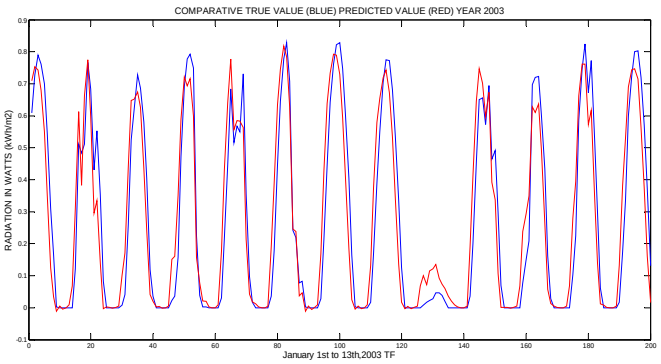

Fig. 4: Comparative with radiation and hour TF using data normalized by 20.

Data fusion of radiation and hour reaches better results than any other combination of phenomena in the study at TF station.

ISBN: 978-1-61804-280-4

4.3 Comparison and discussion

At this point, it is useful to establish a comparison of the results obtained and start a discussion between using normalized data or not.

Table 7 presents a comparison between this study and another previous study with the same configuration of the ANN but using un-normalized data [6].

STUDY	VALUES OF MAE IN KILOWATTS HOUR PER SQUARE METER
[6]	0.04
This work	0.06

Table 6: Comparison of best results between normalized and un-normalized data

At first glance, it is obvious that normalization do not offer any improve on results versus un-normalized data. And if we take a look on results of other normalization blocks like the result of block 2.1 given above, a normalization block does not improve but also worsens previous results like obtained in [6].

The reason for that may reside in the values of solar radiation. If we take a look into any of the figures accompanying results it is easy to see that the variation of any of the samples, for example, *i* sample does not present significant variation versus *i-1* or *i+1* sample.

This linearity of the samples, in conjunction with the low value (near to ±1), it causes that the input data to the ANN does not present significant variations that allows the ANN to recombine the inner weights of samples in order to reduce the gap between predicted value and real value.

It is a valuable goal, in order to apply it to other meteorological parameters. For slow variations of parameter and its value near ±1, it is better to use the non-normalizing data.

5 Conclusions

Once realized the study, it is shown that normalization gives good results when normalization by decades is used and it worsens previous results when normalization between a maximum and minimum values or null mean normalization is applied.

When normalized data and un-normalized data are so close no benefit of normalization is reached.

The system is quick to give a result as shown in the results tables in test times per sample.

6 Acknowledgements

This work has been supported by The Endesa Foundation and The University of Las Palmas Foundation (FULP) under Grant "*Programa Innova Canarias 2020*".

References:

[1] Fang Huang; Shuisen Zhou; Shaosen Zhang; Hongwei Zhang; Weidong Li (September, 2011), "Meteorological Factors–Based Spatio-Temporal Mapping and Predicting Malaria in Central China." *American Journal of Tropical Mededicine and Hygiene*. [On line] 85 (3), pp. 560-567.
Available: http://www.ajtmh.org/content/85/3/560.full.pdf+html [Dec. 5, 2014]

[2] Vasak, M.; Gulin, M.; Vic, J.C.; Nikolic, D.; Pavlovic, T.; Peric, N., "Meteorological and weather forecast data-based prediction of electrical power delivery of a photovoltaic panel in a stochastic framework," *Proceedings of the 34th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO, 2011)*, 2011, pp.733,738, 23-27 May 2011.

[3] Pninit Cohen; Oded Potcher; Andreas Matzarakis (May, 2012), "Daily and seasonal climatic conditions of green urban open spacess in the Mediterranean climate and their impact on human confort," Building and Environment. [On line] 51, pp. 285-295. Available: http://www.sciencedirect.com/science/article/pii/S0360132311004100 [Dec. 5, 2014]

[4] Jianwu Zeng; Wei Qiao (April, 2013), "Short-term solar power prediction using an RBF neural network," *Renewable Energy*. [On line] 52, pp. 118-127.
Available: http://www.sciencedirect.com./science/article/pii/S0960148112006465 [Dec. 5, 2014]

[5] Ma, L.; Li, B.; Yang, Z. B.; Du, J.; Wang, J. (January, 2014), "A New Combination Prediction Model for Short-Term Wind Farm Output Power Based on Meteorological Data Collected by WSN." International Journal of Control and Automation. [On line] 7 (1), pp. 171-180.

Available: http://www.sersc.org/journals/IJCA/vol7_no1/14.pdf [Dec. 5, 2014]

[6] Hernández-Travieso; José G., Travieso,C; Alonso, J.; Dutta, Malay K., "Solar radiation modelling for the estimation of the solar energy generation," *2014 Seventh International Conference on Contemporary Computing (IC3)*, 2014, pp.536,541, 07-09 Aug. 2014.

[7] Haykin, S., *Neural Networks. A Comprehensive Foundation*., Second Edition. Upper Saddle River,NJ: Prentice Hall Intenational,Inc., 1999.

ISBN: 978-1-61804-280-4