Clustering and Prediction of Solar Radiation Daily Patterns

G. Nunnari[1], **and S. Nunnari**[1]
[1]Dipartimento di Ingegneria Elettrica, Elettronica e Informatica, Universitá di Catania, Catania, Italy

Abstract— *This paper addresses the problem of clustering daily patterns of global horizontal solar radiation by using a feature-based approach. A pair of features, referred to as S_r and H_r, representing a measure of the normalized daily solar energy and of the energy fluctuations, respectively, is introduced. Clustering allows to perform some useful statistics at daily scale such as estimating the class weight and persistence. Furthermore, the problem of one-day ahead prediction of the class is addressed by using both hidden Markov models (HMM) and Non-linear Autoregressive (NAR) models. Performances are then assessed in terms of True Positive Rate (TPR) and True Negative Rate (TNR).*

Keywords: Solar Radiation, HMM models, NAR models, Clustering time series, Prediction.

1. Introduction

Solar radiation is a quite irregular kind of time series, due several complex processes such as the clouds cover features, the atmospheric transmittance, the sky turbidity and the pollution level. One problem, dealing with solar radiation time series is that they are scarcely auto-correlated. In particular, at hourly scale the correlation time is about 5 lags, while at daily scale the correlation time is 1 lag (see for instance [1]). Thus while there is some chance to forecast hourly average solar radiation at short term by using auto-regressive models, there are very limited possibilities of forecasting one-day ahead the average value of solar radiation. For this reason, clustering techniques may provide useful tools to get some statistical information at daily scale. Previous work concerning the clustering of solar daily patterns was carried out by [2], based on daily distributions of the clearness index. Classification of solar radiation patterns into three classes, referred to as overcast, partly cloudy and sunny, was proposed by [3], based on the use of 10 min sampling data. Classification into to four classes based on 5m sampling data was proposed by [4].
In this paper we propose a strategy belonging to the class of methods working on features extracted from hourly average samples, which are more widely available from public data base with respect to 10m or 5m average time series. Indeed, different sampling rate requires different algorithms for feature extracting, mainly for measuring the degree of fluctuation of solar radiation at daily scale. The paper is organized as follows: a short description of the considered data set is provided in section 2, while description of features extracted from solar radiation daily patterns is given in section 3. The adopted clustering strategy is described in section 4, while applications, devoted to perform some statistical insights from the time series of classes, is given in section 5. One-day ahead class prediction is addressed in section 6 and, finally, conclusions are traced in section 7.

Table 1: Geographic coordinates of the 12 solar radiation recording stations belonging to the considered data set.

stationID	Lat	Lon	$Elev$	UTC
690140	33.667	-117.733	116	-8
690150	34.3	116.167	626	-8
722020	25.817	-80.3	11	-5
722350	32.317	-90.083	94	-6
722636	36.017	-102.55	1216	-6
723647	35.133	-106.783	1779	-7
724776	38.58	-109.54	1000	-7
725033	40.783	-73.967	40	-5
725090	42.367	-71.017	6	-5
726055	43.083	-70.817	31	-5
726130	44.267	-71.3	1910	-5
726590	45.45	-98.417	398	-6

2. The considered solar radiation data set

The data set considered in this paper consists of hourly average time series recorded at twelve stations stored in the USA National Solar Radiation Database managed by the NREL (National Renewable Energy Laboratory). Data of this database was recorded from 1999 to 2005 and can be freely download from ftp://ftp.ncdc.noaa.gov/pub/data/nsrdb-solar/. The twelve stations were selected based on two criteria: the quality of time series and the need to ensure the necessary diversification of meteo-climatic conditions. The twelve selected stations are listed in Table (1). More detailed information about these and others recording stations of the National Solar Radiation Database can be found in [5].

3. Two features of solar radiation

In order to choose a limited number of representative features of the solar radiation time series, it is quite natural to choose a pair that can represent the quantity of solar radiation recorded during a day and the level of fluctuation. Evidently, while the first feature is directly connected with the quantity of electrical energy per day that is expected to

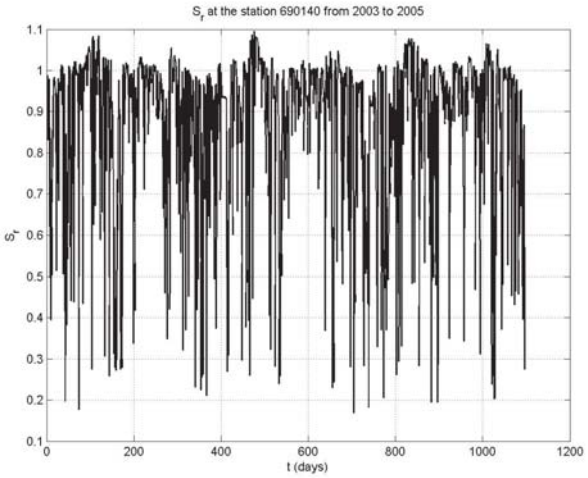

Fig. 1: Daily values of the S_r feature computed for the station ID690140 from 2003 to 2005.

produce, the second gives a measure of its intermittency. The formal definition of the proposed features is given below.

3.1 The solar radiation ratio S_r

The S_r feature is formally represented by expression (1)

$$S_r(t) = \frac{S_{pat}(t)}{S_{csk}(t)} \tag{1}$$

where $S_{pat}(t)$ and $S_{csk}(t)$ represent the area under the true and the global horizontal solar irradiation daily patterns in clear sky conditions, respectively, and t is the time expressed in days. The $S_{csk}(t)$ can be computed referring to one of several existing clear sky models, such as the Ineichen and Perez model [6], [7]. The Matlab code to implement such a model is part of the SNL_PVLib Toolbox, developed in the framework of the Sandia National Labs PV Modeling Collaborative (PVMC) platform. Of course the S_r feature is always positive but can be greater or less than 1: in a day featured by favorable weather conditions (e.g. absence of cloud cover and good atmospheric transmittance) S_r can be slightly greater than 1; conversely under thick cloud cover and adverse propagation conditions it may be significantly less than 1. As an example, the behavior of $S_r(t)$ computed for the station ID690140 from 2003 to 2005 is reported in Figure (1).

3.2 The Hurst exponent ratio

The Hurst exponent ratio H_r is formally represented by expression (2)

$$H_r(t) = \frac{H_{pat}(t)}{H_{csk}(t)} \tag{2}$$

where $H_{pat}(t)$ and $H_{csk}(t)$ are the Hurst exponent of the true and clear sky solar radiation patterns, respectively, at the generic day t. The rationale for this definition is that while smooth curves, such as solar radiation in clear sky conditions, gives H_r close or slightly greater than 1, the presence of fluctuations generates H_r less than 1. For the purposes of this paper the Hurst exponent was computed by using the Geweke-Porter-Hudak (GPH) algorithm, first described by [8]. The reason for using this algorithm, instead of the more popular R/S and DFA algorithms, is that it is considered more appropriate when patterns, as in this paper, are represented by a small number of samples. Indeed, the GPH algorithm is based on the slope of the spectral density function. An example of H_r index, computed for the station ID690140 during three years, is shown in Figure (2).

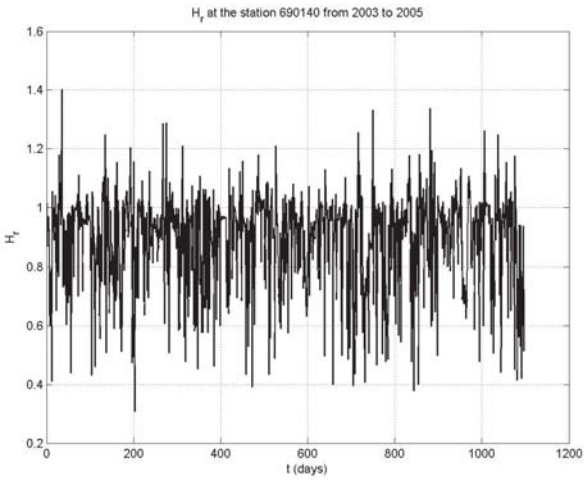

Fig. 2: H_r daily values computed for the station ID690140 from 2003 to 2005.

3.3 Some properties of the S_r and H_r features

The extremely scattered behavior of these features can be objectively evaluated by computing the mutual information of the individual S_r and H_r, as shown in Figure (3). Indeed, the mutual information suddenly decays after just 1 lag. This result implies that one-day ahead prediction of the class is extremely difficult by using auto-regressive models. Therefore, clustering approaches can be useful to extract, at least, some statistical information at daily scale. It is worth noting that S_r and H_r are not really independent features, as shown by the cross-correlation function reported in Figure (4). In more detail, they are positive-correlated, thus meaning that high H_r correspond to high S_r, as mentioned in section 3.2. The role of S_r and H_r to classify solar radiation daily patterns is discussed in the next section.

4. Clustering solar radiation features

The general goal of clustering is to identify possible structures in an unlabeled data set, in such a way that data is objectively organized in homogeneous groups. The heart of

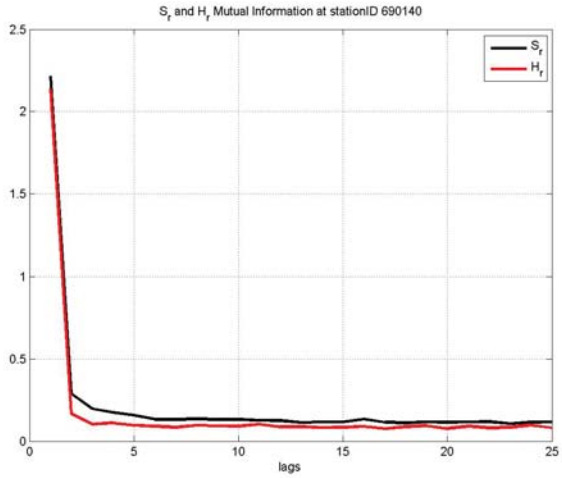

Fig. 3: Mutual information of Sr and H_r computed for the station ID690140.

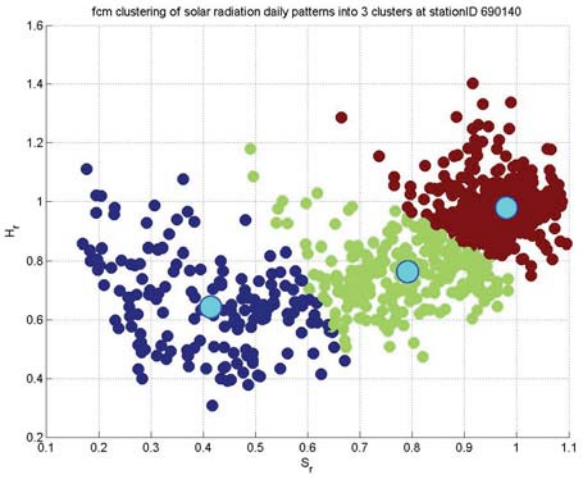

Fig. 5: Pattern distribution and cluster centers computed for the station ID690140.

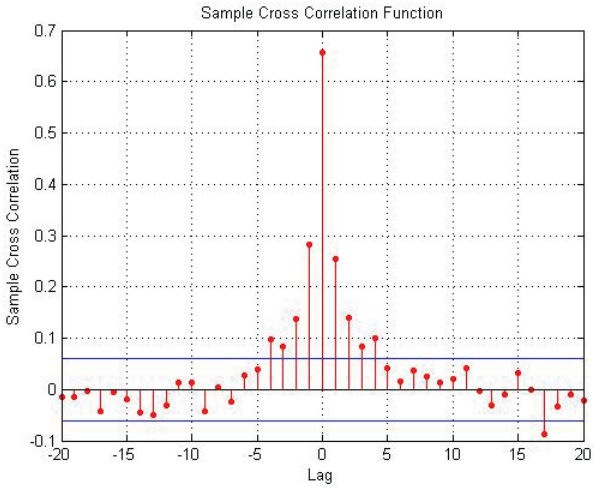

Fig. 4: Cross-correlation between Sr and H_r computed for the station ID690140.

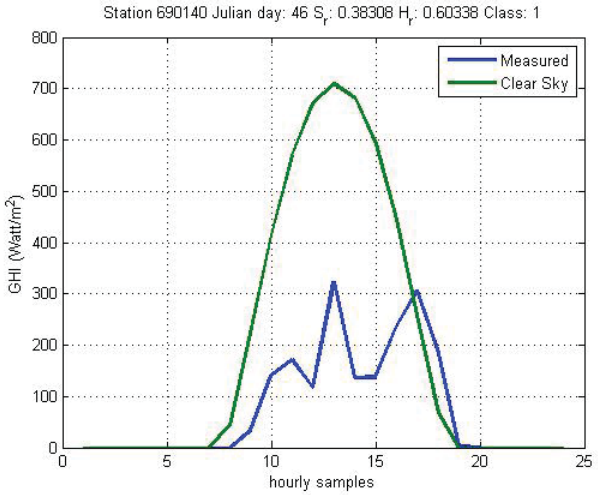

Fig. 6: Example of class C_1 pattern at the station ID690140.

any clustering approach, regardless of whether the problem is to classify static data or time series, is a clustering algorithm. The most important existing algorithms are usually grouped into four groups: exclusive clustering, overlapping clustering hierarchical clustering and probabilistic clustering. For the purposes of this work, since it is not realistic to perform several clustering levels, hierarchical clustering is not appropriate. In this paper, we have considered the fuzzy-c means (*fcm*) algorithm, which is simple to implement, allows overlapping clustering and generate only one level of clusters. Results of clustering into 3 classes the features computed for the station ID690140 is shown in Figure (5). As it is possible to see the *fcm* algorithm essentially distributes the cluster centers for increasing values of S_r which thus play the role of primary feature. Furthermore, as already observed, patterns characterized by high S_r are also featured by high values of H_r.

Representative patterns, in a 3-class framework, are shown in Figures (6), (7) and (8), respectively. In order to assess the consistency of clustering by using the *fcm* algorithm, we have computed the silhouette, which for the station ID690140 looks as in Figure (9). The silhouette $S(i)$ is a measure, ranging from -1 to 1, of how well the i_{th} pattern lies within its cluster: $S(i)$ close to 1 means that the corresponding pattern is appropriately clustered; on the contrary, if $S(i)$ is close to -1, then the i_{th} pattern would be more appropriate if it was clustered in its neighboring cluster; finally $S(i)$ near to zero means that the pattern is on the border of two natural clusters. As it is possible

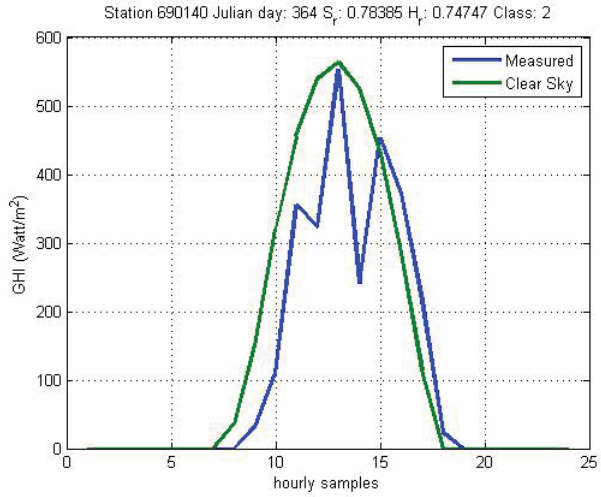

Fig. 7: Example of class C_2 pattern at the station ID690140.

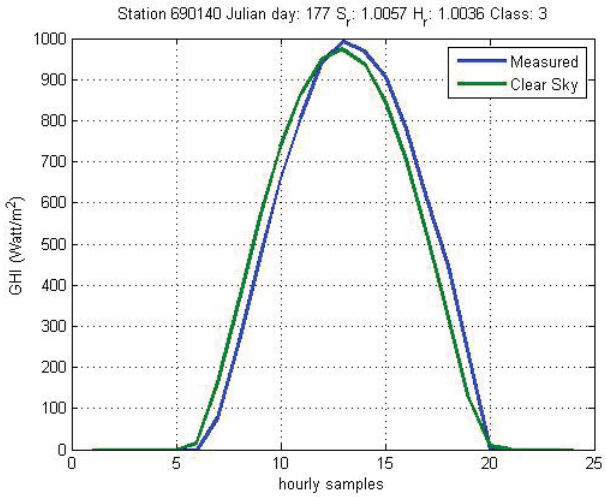

Fig. 8: Example of class C_3 pattern at the station ID690140.

to appreciate from Figure (9), only a limited number of events of classes C_1 and C_2 are characterized by negative silhouette, thus assessing the goodness of the clustering.

It is to be stressed that the coordinates of the cluster centers are depending on the recording site, as shown in Figure (10), where the centers computed at the 12 considered stations, within a 3-class framework, are reported.

5. Some applications

Once classes have been attributed to daily wind speed time series, some useful statistics can be carried out, such as, for instance, computing the weight of a class and the persistence of patterns in a class.

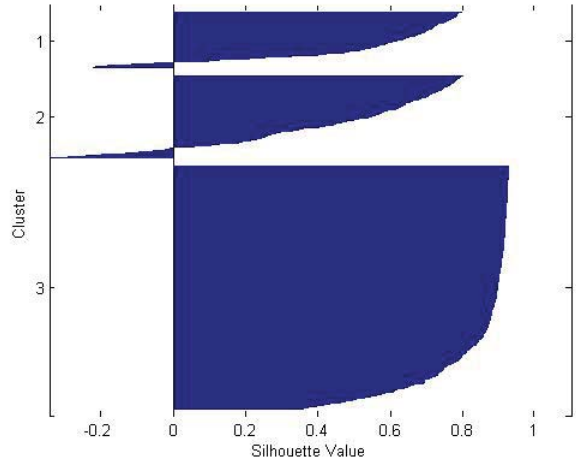

Fig. 9: Silhouette computed clustering patterns of the station ID690140 into 3 classes by the *fcm* algorithm.

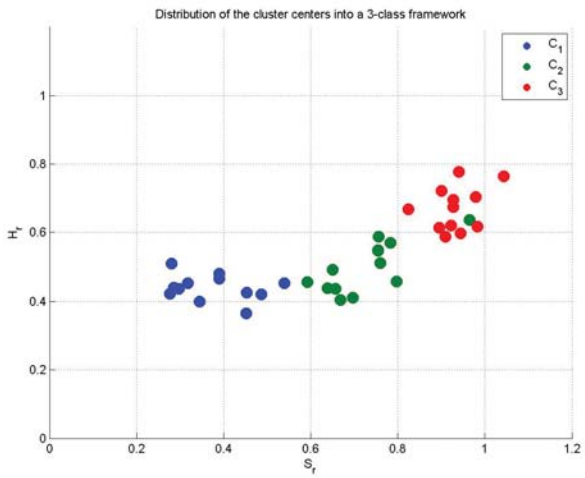

Fig. 10: Distribution of the cluster centers at 12 stations into a 3-class framework.

5.1 Weight of a class

The weight W_i of a class C_i is formally defined as in (3)

$$W_i\% = \frac{n_i}{\sum_{i=1}^{n_c} n_i} 100 \tag{3}$$

where n_i and n_c are the number of patterns in class C_i and the number of classes, respectively. The weights computed for each station of the considered data set are reported in Table (2), As it is possible to see, the weights are heavily affected by the different meteo-climatic conditions in which operates each individual recording station. In particular, Table (2) shows that class C_3 exhibits on average the highest weight.

Table 2: Weight in percent of the four classes at 12 stations.

stationID	$W_1\%$	$W_2\%$	$W_3\%$
690140	14.69	21.53	63.78
690150	11.13	30.38	58.49
722020	23.18	34.03	42.79
722350	21.72	24.45	53.83
722636	15.51	24.73	59.76
723647	13.78	24.27	61.95
724776	18.16	29.47	52.37
725033	27.46	30.57	41.97
725090	26.55	27.01	46.44
726055	27.01	27.83	45.16
726130	31.20	37.04	31.75
726590	24.73	29.20	46.08
Average W	**21.26**	**28.38**	**50.36**

5.2 Estimating the persistence

A useful statistic is that of estimating the persistence, defined as the number of episodes in a year in which a daily pattern persists in the same class for at least p consecutive days. An example of this kind of statistic is given in Figure (11). The Figure refers to the persistence of patterns at the

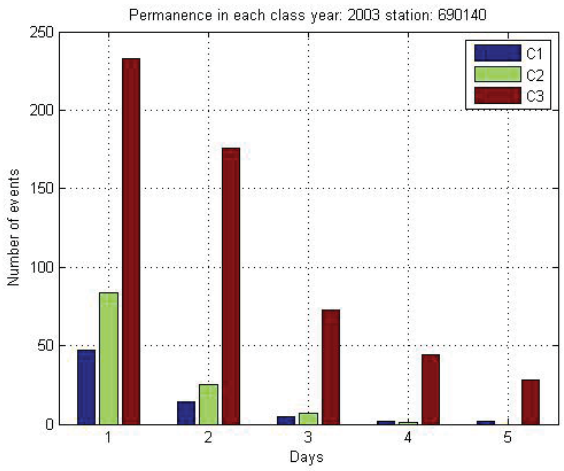

Fig. 11: Persistence in the same class at the station ID690140 during 2003.

station ID690140 during 2003 and show, for instance, that 176 events lasted in class C_3 at least two days, 73 at least 3 days, 44 of at least 4 days and 28 at least 5 days. Instead, for patterns of class C_2, only 25 lasted at least 2 days, 7 at least 3 days and 1 at least 4 days. This kind of information could be useful to a solar plant manager in the absence of reliable predictions at daily scale.

6. Predicting one-day ahead the class

In this section we report results concerning some attempts to predict one-day ahead the time series of solar radiation class, obtained as described in the previous section 4. Of course, the prediction problem is as difficult as larger is the number of considered classes. Here we report results concerning prediction in 2 and 3 class frameworks, since as explained in section 6.4, at the present stage of this work, results are not reliable for larger number of classes. As concerning the prediction approaches we have considered HMM and NAR models.

6.1 Predicting by using HMM models

A HMM [15] is a modeling approach in which we observe a sequence of emissions, but we do not know the sequence of states the model went through to generate the emissions. Thus, in general a HMM model is characterized by two matrices, referred to as the transition matrix and the emission matrices, respectively. Prediction by using this kind of models requires that the emission matrix must be transformed into the most probable state path by using one of several available algorithms, such as the popular Viterbi algorithm [16].

6.2 Predicting by using NAR model

The NAR-based model consists of two main steps. In the first step, one-day-ahead prediction of $\hat{S}_r(t+1)$ and $\hat{H}_r(t+1)$) of the two features $S_r(t)$ and $H_r(t)$ are performed by using independent models of the form (4).

$$y(t+1) = f(y(t), y(t-1), ..., y(t-d+1)) \qquad (4)$$

In expression (4) d (dimension) is the number of considered past values and f a non-linear unknown map, here identified by using a neural network approach. In the second step, the predicted class $c(t+1)$ is obtained by using a neural network classifier, previously trained to assign a class to a predicted pair of features ($\hat{S}_r(t+1)$,$\hat{H}_r(t+1)$)).

6.3 Performance indices

In order to objectively asses to what extent a predicted time series of classes is close to the true one it is possible to use several performance indices. In particular, in this paper we have considered the TPR (True Predicted Rate) and the TNR (True Negative Rate), defined as follows:

$$TPR(i) = \frac{TP(i)}{TP(i)+FN(i)} \qquad (5)$$

$$TNR(i) = \frac{TN(i)}{TN(i)+FP(i)} \qquad (6)$$

where $TP(i)$ and $FN(i)$ is the number of true positive and false positive patterns, respectively, attributed by the model to the class C_i and i is the class index. The sum $P(i) = TP(i) + FN(i)$ is, of course the total number of patterns attributed by the model to the class C_i. Similarly, $TN(i)$ is the number of patterns which are correctly identified as not belonging to the class C_i and $FP(i)$ is the number of false positives attributed by the model to the class C_i. The sum $N(i) = TN(i) + FP(i)$ is the total number of patterns recognized by the model as not belonging to the

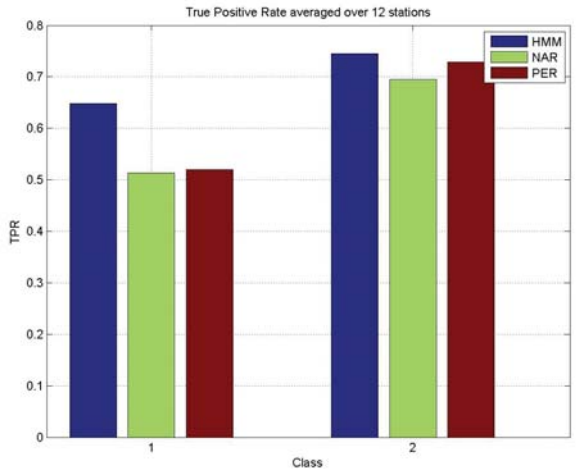

Fig. 12: TPR for the 2-class framework averaged over twelve stations.

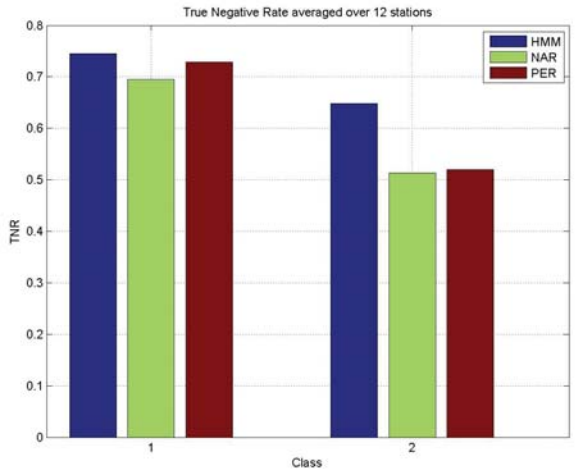

Fig. 13: TNR for the 2-class framework averaged over twelve stations.

class C_i. Clearly, a good predictor would be characterized by values of TPR and TNR both close to 1. Performances of the HMM and NAR models were also compared with the simple persistent model (7), often considered as a low reference model.

$$c(t+1) = c(t) \tag{7}$$

6.4 Numerical results

Results described in this section was obtained by using for each considered station hourly average solar radiation time series, recorded from 2003 to 2005. For each station, two years of data (2003 and 2004) was considered to identify the HMM and NAR prediction models, while the remaining year 2005 was considered to test the models. In order to generalize the results, the performance indices shown in this section were averaged over the whole set of considered stations.

6.4.1 Prediction into a 2-class framework

We start the description with the simplest case, i.e. the prediction in a 2-class framework. The TPR and TNR, averaged over twelve stations are reported in Figure (12) and (13), respectively. It is possible to see that the HMM prediction model outperform, both the NAR and the persistent models, exhibiting an average TPR of about 0.65 and 0.75 for the classes C_1 and C_2. Similarly, the TNR is about 0.75 and 0.65 for the classes C_1 and C_2, respectively. A detail of the TPR and TNR of the HMM model obtained for each individual station is reported in Figure (14) and (15), respectively, which show that the model performances, are significantly affected by the recording site. In this Figure, different colors refer to different stations; the order of the station is the same as in the first column of Table (2). In the background of the Figure the TPR averaged over the twelve stations is given.

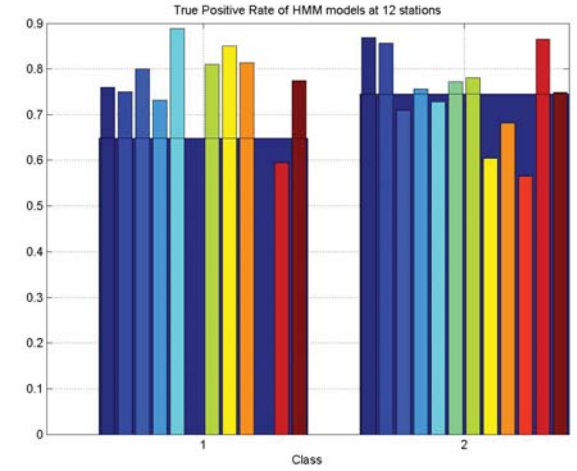

Fig. 14: TPR of the HMM model for each of the 12 station, in the 2-class framework.

6.4.2 Prediction into a 3-class framework

As concerning the models performances into a 3-class framework, the average TPR and TNR are reported in Figure (16) and (17), respectively. In particular, Figure (16) shows that despite the considered models are capable of predicting an acceptable proportion of patterns belonging to the class C_3, they are not as good to correctly predict patterns belonging to the classes C_1 and $C2$, since the corresponding TPR is below 0.5. That notwithstanding, Figure (17) shows that all models have a good ability to correctly recognize patterns that do not belong to a particular class. Furthermore, for the 3-class framework, there is not a clear prevalence of a model with respect to the other.

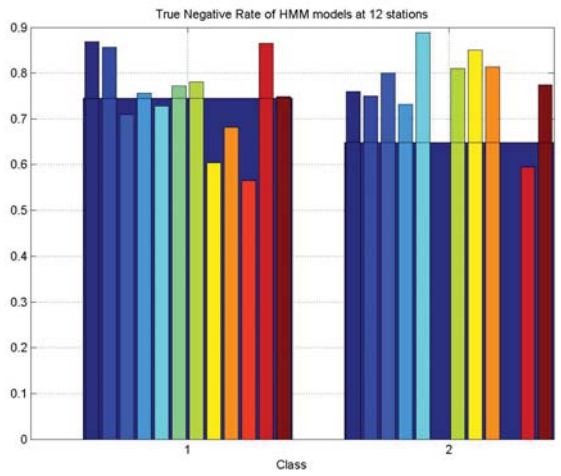

Fig. 15: TNR of the HMM model for each of the 12 station, in the 2-class framework.

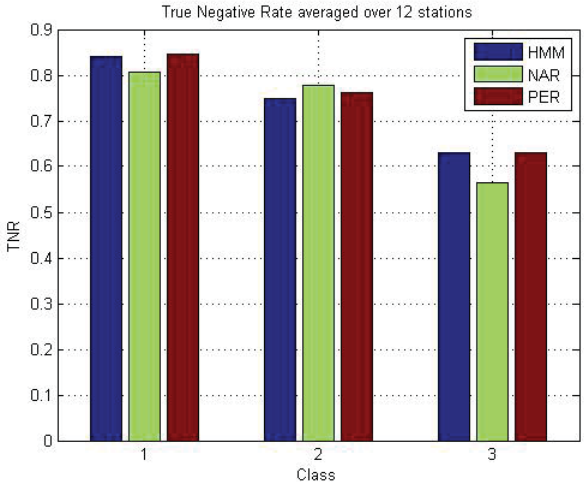

Fig. 17: TNR for the 3-class framework averaged over twelve stations.

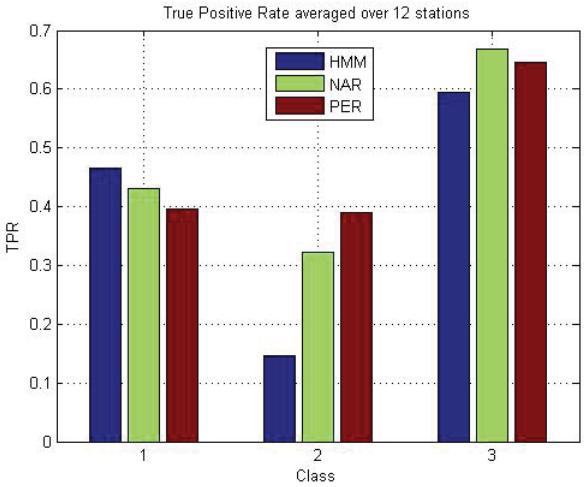

Fig. 16: TPR for the 3-class framework averaged over twelve stations.

7. Conclusions

In this paper a feature based strategy to cluster solar radiation daily patterns was presented, which allows associating to the original solar radiation time series a time series of classes. This allows to perform some useful statistics which would be otherwise not possible to make, such as the class weight and the persistence analysis. Furthermore, the paper has addressed the problem of one-day ahead prediction of the class. To this purpose two different approaches were considered, namely the HMM and the NAR approach. Results show that, at the present stage of this work, the prediction models are effective for a 2-class framework only. Work is still in progress to improve these results.

8. Acknowledgements

The research was supported by the Universitá di Catania under the grant FIR 2014.

References

[1] L. Fortuna, G. Nunnari, S. Nunnari, Nonlinear modeling of solar radiation and wind speed time series, SpringerBrief in Energy, ISBN 978-3-319-38763-5.
[2] T. Soubdhan, R. Emilion, R. Calif, Classification of daily solar radiation distributions using a mixture of dirichlet distributions, Solar Energy 83 (2009) 1056–1063. doi:10.1016/j.chaos.2008.07.020.
[3] M. Nijhuis, B.G. Rawn, M. Gibescu, Classification technique to quantify the significance of partly cloudy conditions for reserve requirements due to photovoltaic plants, Proceedings of the 2011 IEEE Trondheim PowerTech Conference, 2011.
[4] L. Fortuna, G. Nunnari, S. Nunnari, A new fine-grained classification strategy for solar daily radiation patterns,Pattern Recognition Letters, 2016, DOI: 10.1016/j.patrec.2016.03.019.
[5] S. Wilcox, National Solar Radiation Database 1991–2010 Update - Users Manual, Technical Report NREL/TP-5500-54824, 1–479, 2012. ftp://ftp.ncdc.noaa.gov/pub/data/nsrdb-solar/documentation-2010/.
[6] P. Ineichen and R. Perez, A New airmass independent formulation for the Linke turbidity coefficient,Physica A,2002,73,151–157.
[7] R. Perez, A New Operational Model for Satellite-Derived Irradiances-Description and Validation, Solar Energy, 2002,73, 207–317.
[8] J. Geweke, S. Porter-Hudak, J. Time Series Analysis 4 (1983) 221.
[9] R. Weron, Estimating long range dependence finite sample properties and confidence intervals, Physica A 312 (2002) 285–299.
[10] T. Kohonen, Self-Organizing Maps, 1995.
[11] T. W. Liao, Clustering of time series data - a survey, Pattern Recognition 38 (2005) 1857–874.
[12] L. A. Zadeh, Fuzzy sets, Information and Control 8 (1965) 338–353.
[13] T. Kohonen, Self-Organizing Maps, 1995.
[14] T. W. Liao, Clustering of time series data - a survey, Pattern Recognition 38 (2005) 1857–874.
[15] L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77 (2), 257Ű286, 1989.
[16] A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE Transactions on Information Theory 13 (2), 260Ű269. doi:10.1109/TIT.1967.1054010.