

Finding relevant multivariate models for multi-plant photovoltaic energy forecasting

Youssef Hmamouche*, Piotr Przymus[†], Lotfi Lakhali* and Alain Casali*

LIF - CNRS UMR 7279, Aix Marseille University, Marseille, France,

*firstname.lastname@lif.univ-amu.fr, [†]piotr@przymus.org

Abstract. Forecasting the photovoltaic energy power is useful for optimizing and controlling the system. It aims to predict the power production based on internal and external variables. This problem is very similar to the one of multiple time series forecasting problem. With the presence of multiple predictor variables, not all of them will equally contribute to the prediction. The goal is, given a set of predictors, to find what is the best / most accurate subset (s) leading to the best forecast. In this work, we present a feature selection and model matching framework. The idea is that we try to find the optimal combination of forecasting model with the most relevant features for given variable. We use a variety of causality based selection approaches and dimension reduction techniques. The experiments are conducted on real data and the results advocate the usefulness of the proposed approach.

Keywords: Time Series; Prediction; Data Mining; Ensemble Selection.

1 Introduction

Time series forecasting is an important tool aiming to predict the evolution of time series over time based on the existing history. It has many applications, for example, in finance, neuroscience, industrial optimization ... this field is considered as an essential part of business intelligence systems. It delivers crucial information that can improve the decision making processes, by anticipating systems behavior, e.g., energy consumption or production. Forecasting photovoltaics (PV) Energy Production has gained attention with the growing of interest in using PV as source of renewable energy. Forecasting the production of such systems has a direct impact on trading and controlling the used energy.

In general, the PV energy can be measured as time series variables that can change according to the system state and external conditions, like temperature and the weather conditions. The simplest approach would be to use univariate forecasting model for power generation time series. Several models can be used in this context, for example the auto-regressive models, e.g. AR or ARIMA [1]. However this option have some drawbacks: it does not include crucial informations provided by other variables. In this case, it is worth to exploit this extra information from other variables using multivariate models. One approach would

be to use all available variables, but this (i) incorporate some irrelevant variables, and thus decrease the forecast accuracy [2], and (ii) use too much memory. Such situation can be improved by extracting only the most relevant variables. This rises some interesting challenges for multivariate time series forecasting. The organization of the paper is as follow. In the first section, we present and discuss some works related to the problem addressed. In the Section, 3 we detail the proposed method. In Section 4, we describe the forecasting process and the methodology used to perform the experiments. In Section 5, we show and discuss the results. And in the last section, we summarize our approach.

2 Related Work

In the literature, many approaches was proposed to handle the problem of forecasting PV energy production. In [3], the paper deals with multi-plant PV energy production forecasting. A comparison between artificial neural networks, regression trees, and spatio-temporal auto-correlation based methods was experimented. The authors show that regression trees provide better results than artificial neural networks (ANNs). In [4], ANNs are used to forecast PV energy production, taking advantage from there ability to learn the changes. To improve the forecasts, multiple predictor variables that may influence the energy production were used based on internal and external factors. The same problematic was investigated in [5]. An hybrid approach was used, by adding basic physical constraints of the PV plant to the input of an ANN. The results show an improvement of prediction accuracy compared to model without those constraints. More works on photovoltaic power forecasting approaches can be found in [6].

We argue that the problem of PV energy forecasting can be modelled as multivariate time series prediction. In the following we reformulate this problem and discuss the main approaches used to address it. Consider a set of predictor time series $X = [x_1, \dots, x_k]$ and a target variable y , with n observations.

There are multiple strategies to predict y using X . One way consists in using models that exploit the precedent values of y and X , e.g., the vector autoregressive models [7]. In this work, we focus on the prediction models that predict Y at time t based on values of variables of X at the same time t . Therefore, the general model can expressed as follow, $y(t) = f(X_1(t) + \dots + X_k(t)) + \epsilon(t)$.

Linear models suppose that y can be expressed as linear combination of X , i.e., $y(t) = \beta_0 + \sum_{i=1}^k \beta_i X_i(t) + \epsilon(t)$, where $\epsilon(t)$ is the error term, and $\beta = [\beta_0, \beta_1, \dots, \beta_k]'$ is the vector parameter of the model. The estimation of these parameters can be performed via different methods. The most common one is the Least Square technique, which consists on minimizing the sum of squared errors, and the resolution is performed based on straightforward derivation.

Shrinkage methods aim to minimize the impact of irrelevant variables by setting the coefficients close to zero. These technique is practical where the number of predictors is large and the classical resolution is not possible due to matrix operations constraints. For instance, the Ridge regression method proposed in [8], minimizes the term $\sum_{t=1}^n (y(t) - \beta_0 - \sum_{i=1}^k \beta_i X_i(t))^2 + \lambda \sum_{j=1}^k \beta_j^2$, where

$\lambda \sum_{j=1}^k \beta_j^2$ is the shrinkage penalty. This mechanism results in shrinking estimated coefficients towards zero. The Least Absolute Shrinkage and Selection Operator (Lasso) method is similar to the Ridge regression, but it uses $\lambda \sum_{j=1}^k |\beta_j|$ as a shrinkage penalty term, in order to force the coefficient of unimportant variables to be equal to zero.

ANNs use generally a non-linear function (a network of nodes, where each one pass the signal using weight and eventually an activation function). They are characterized by the ability of modelling the dynamic dependencies between variables and learning from the precedent information passed through the network. By considering the prediction training step as a supervised problem, the main algorithms used to calibrate the coefficients of the network are based on the back-propagation of the errors using for instance gradient descent or stochastic gradient descent algorithms [9], [10].

To handle the problem of selecting the most important predictors in a multivariate prediction model, different approaches based on dimension reduction and feature selection techniques were proposed in the literature. In [11], a comparison of five dimensionality reduction and feature selection methods (t-test and correlation based method (ranking technique), step-wise regression, principle component analysis and factor analysis) is performed as a pre-processing step to improve the forecast accuracy. Also in [12], the authors combine multiple dimension reduction methods based on Principal Component Analysis (PCA), Genetic Algorithms (GA) and decision trees (CART), to improve the multivariate prediction models with all existing variables. In [13], a feature selection algorithm based on causality is proposed for stock prediction modeling. To avoid the main problem of correlation, i.e., it cannot distinguish direct influences from indirect ones. The authors select variables based on causality. This method was compared with PCA, decision trees and LASSO. In [14], an overview of methods that uses principal components approaches for regression. And a sufficient method for regression with many predictors was proposed.

3 The Proposed Feature Selection Method

In this section, we expose our proposed method. Let us consider a target variable y and a set of predictors P . The goal is extract the relevant variables from P , i.e a subest of P , based on the notion of causality, that will be used in a model to forecast y . Our approach consists of three steps. First, we calculate the graph of causalities, then we reduce it by eliminating dependencies using a simple transitive reduction technique. Finally, we rank them regards with the causality on the target variable.

To compute causality, we use two measures: (i) the Granger causality [15] and (ii) the Transfer entropy [16]. They are characterized by the property of modeling non-symmetric relationships between variables. In other sense, they detect which variable has a direct impact on the other one.

Let us consider two univariate time series x_t, y_t . The Granger causality assumes that x_t causes y_t if it contains helpful information to predict y_t . The as-

sociated test estimates causality using the Vector Auto-Regressive model. Two models are computed, one using just the values of the target variables, and the second using the target and the predictor variables. Then a difference between with those two models is evaluated using the F-test. In the other hand, Transfer Entropy has similar idea in evaluating the behavior of the target variable by using itself and the predictor variable, but it is based on information theory. Let us underline that Granger causality is based on a prediction model while Transfer entropy is based on information theory. It has been shown in [17] that they are equivalent only for variables following a normal distribution.

The goal of the proposed method is simple, extracting variables by ranking them according to the causality. However, selecting them directly based on such non-symmetric measure leads to the problem of dependencies between variables. In other words, it is possible to select a set of variables in which each one cause the other, or even they can be duplicated (they could contain the same information used to predict the target). Hence, a diversification can improve the selection task. In this case, applying the transitive reduction algorithm seems natural as a processing step. We summarize in Algorithm 1 our method. A short version is provided where we suppose that the causality graph is input of the algorithm. The following notation are adopted: $x \rightarrow y$ expresses the fact that x causes y , and $\text{causality}(x \rightarrow y)$ is the value of this causality.

Algorithm 1: Transitive Reduction on Causality Graph (TRCG)

Input: The causality graph G , the target variable y , the reduction size k

Output: S : Set of predictor variables of y .

```

/* Eliminating dependencies with regard to the target variable */
1: for all node  $ts_1 \in G.nodes \setminus \{y\}$  do
2:   for all node  $ts_2$  in  $G.nodes \setminus \{ts_1, y\}$  do
3:     if  $ts_1 \rightarrow ts_2, ts_2 \rightarrow y$  and  $ts_1 \rightarrow y$  then
4:       Remove edge between  $ts_1$  and  $y$ 
/* Selecting top  $k$  variable (nodes of  $G$ ) that cause  $y$  */
5:  $P = \{ts \in G.nodes, ts \rightarrow y\}$ 
6:  $P_s = P.sort$  (key= $\lambda x: \text{causality}(x \rightarrow y)$ )
7:  $S = top_k(P_s)$ 
8: return  $S$ 

```

4 Methodology

The data sets experimented are hourly multiple time series (from hour 2 to 20 each day), representing 3 PV plants, spanning a period of 12 months (year 2012). The goal is to predict 3 months of the production variable, from January to March 2013 (where the values of target variables are not known), based on internal factors (temperature and irradiance), and external factors (cloudcover,

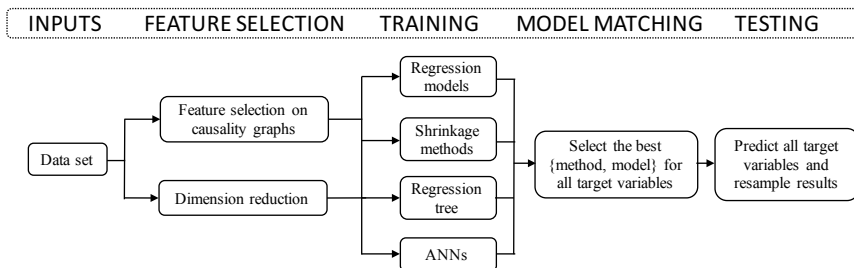


Fig. 1: The used forecasting process

dewpoint, humidity, pressure, temperature, windbearing, windspeed). The data are organized in a way to predict each hour separately, i.e., for each plant, we have 19 target variables to predict.

The methodology adopted is based on model selection. First, a benchmark experiment is performed on training data (year 2012) using cross-validation with 8 experiments by predicting 3 months in each experiment. We execute all the models on the subsets generated by all the methods. Then we select for each target variable a pair {method, model} that will be used in testing step.

In the reduction step, we use two existing methods, the Random Walk with Restart on Granger causality graphs GRWR and on Transfer entropy graph TRWR [18], and the PCA method. Two versions of each method proposed in 1, TTRCG and GTRCG, is using either transfer entropy or Granger causality for the causality measure. The forecasted models used can be classified in four main types:

- Regression models: Linear regression, RANSAC Regressor (RR), Orthogonal Matching Pursuit (OMP), Theil Sen Regressor (TSR), Hibber Regressor (HB).
- Regression models with shrinkage representation: Ridge, Bayesian Ridge, SVM, Lasso.
- Decision trees: Decision Tree Regressor (DTR), Gradient Boosting Regressor (GBR)
- ANNs: a simple multilayer perceptron neural network (MLP), using one hidden layer and a stochastic gradient descent algorithm to update the parameters of the network.

5 Results and Discussions

In this section we present obtained results, and we provide discussion. As we described in the previous section, we used 3 heuristics (PCA, RWR and TRCG) in the training step. In testing step, we also used the brute force feature selection approach that compute all the possible subset for small number of the fastest prediction models. This allowed us to improve a few of models that were previously pre selected using heuristic approaches. We obtained RMSE= 0.177 for

10% of testing data and 0.253 for all testing data. In the following we present the result of the ensemble selection approach obtained in training step, i.e., with heuristics methods. We focus on the results with heuristic methods, as they can be applied for large-scale data sets.

Table 1: Results of model selection step

Hours	Model selection for all PV plants					
	id ₁		id ₂		id ₃	
	Method	Model	Method	Model	Method	Model
2	GTRCG	GBR	GTRCG	MLP	GRWR	GBR
3	TRWR	GBR	TRWR	MLP	GRWR	GBR
4	GRWR	GBR	GRWR	MLP	TTRCG	GBR
5	TRWR	Lasso	TRWR	HB	TTRCG	Lasso
6	TRWR	HB	TTRCG	HB	GRWR	HB
7	GRWR	MLP	GRWR	MLP	TTRCG	MLP
8	TRWR	TSR	GRWR	TSR	TRWR	HB
9	TRWR	GBR	TTRCG	HB	TRWR	Lasso
10	GTRCG	GBR	GRWR	OMP	TRWR	Lasso
11	GTRCG	MLP	GTRCG	MLP	TRWR	TSR
12	TRWR	TSR	GRWR	OMP	TRWR	TSR
13	TRWR	HB	TTRCG	OMP	TRWR	HB
14	TRWR	HB	TRWR	TSR	TRWR	HB
15	TRWR	HB	TRWR	HB	TTRCG	HB
16	GRWR	HB	TTRCG	HB	TRWR	HB
17	TRWR	HB	TRWR	TSR	GTRCG	GBR
18	TTRCG	GBR	TTRCG	MLP	TTRCG	MLP
19	TRWR	GBR	TRWR	MLP	GTRCG	GBR
20	GRWR	GBR	GRWR	MLP	GTRCG	GBR

Table 1 shows that causal-based feature selection methods outperform the PCA based approaches. Especially approaches based on RWR on the graph of causalities [18] and the newly proposed algorithm are the most competitive. But the general picture is that there is no model that gets best results in all cases.

In Figure 2, the forecast accuracy based on RMSE show that plants id₁ and id₂ are quite similar. Both in terms of characteristics and selected models (1).

In the same figure, relative RMSE are showed in the last three plots. We remark that there exist some hours at the beginning and the end of the day (from 2 to 5 and from 18 to 20) when the energy production is weak and they are very hard to predict. Unfortunately this decrease the global forecast accuracy. As a side remark, the selected prediction models for these hours are the MLP and Gradient Boosting models for all the plants. Which means that when the energy production is low, the data is prone to have some outliers and missing values.

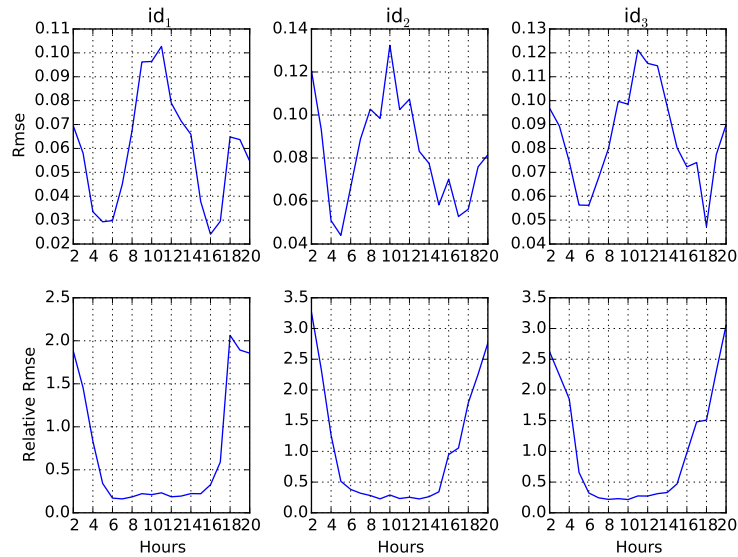


Fig. 2: Forecast accuracy analysis using RMSE

6 Conclusion

In this paper we investigated the multi-plant PV energy forecasting task. We presented an feature selection and model matching framework. The idea is that, for a given variable, we can use heuristics to find the optimal combination of a forecasting model with the most relevant features. Our matching approach is a two step process: *(i)* we use an algorithm that picks optimal subset of features (or combines the features), and *(ii)* we evaluate the selection on various prediction models, like regression, decision trees or artificial neural network models. Finally we select models that perform the best. The second contribution is a new feature selection algorithm, which uses the transitive reduction algorithm on the graph of causalities. The results show the utility of using different feature selection methods and prediction models. However, the forecast accuracy analysis using relative mean squared errors shows some difficulties to give good predictions in a decent time, especially when the energy production is low, which decrease the global performance.

References

1. Box, G.: Box and Jenkins: Time Series Analysis, Forecasting and Control. In: A Very British Affair. Palgrave Advanced Texts in Econometrics. Palgrave Macmillan UK (2013) 161–215
2. Stock, J.H., Watson, M.W.: Chapter 10 Forecasting with Many Predictors. In G. Elliott, C.W.J.G., Timmermann, A., eds.: Handbook of Economic Forecasting, Volume 1. Elsevier (2006) 515–554

3. Ceci, M., Corizzo, R., Fumarola, F., Malerba, D., Rashkovska, A.: Predictive Modeling of PV Energy Production: How to Set Up the Learning Task for a Better Prediction? *IEEE Transactions on Industrial Informatics* **13**(3) (June 2017) 956–966
4. Dumitru, C.D., Gligor, A., Enachescu, C.: Solar Photovoltaic Energy Production Forecast Using Neural Networks. *Procedia Technology* **22** (January 2016) 808–815
5. Gandelli, A., Grimaccia, F., Leva, S., Mussetta, M., Ogliari, E.: Hybrid model analysis and validation for PV energy production forecasting. In: 2014 International Joint Conference on Neural Networks (IJCNN). (July 2014) 1957–1962
6. Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de Pison, F.J., Antonanzas-Torres, F.: Review of photovoltaic power forecasting. *Solar Energy* **136** (October 2016) 78–111
7. Johansen, S.: Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica* **59**(6) (1991) 1551–1580
8. Hoerl, A.E., Kennard, R.W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**(1) (1970) 55–67
9. Zhang, T.: Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms. In: Proceedings of the Twenty-First International Conference on Machine Learning. ICML '04, New York, NY, USA, ACM (2004) 116–
10. Bottou, L.: Stochastic Gradient Descent Tricks. In: *Neural Networks: Tricks of the Trade*. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg (2012) 421–436
11. Tsai, C.F.: Feature selection in bankruptcy prediction. *Knowledge-Based Systems* **22**(2) (March 2009) 120–127
12. Tsai, C.F., Hsiao, Y.C.: Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems* **50**(1) (December 2010) 258–269
13. Zhang, X., Hu, Y., Xie, K., Wang, S., Ngai, E.W.T., Liu, M.: A causal feature selection algorithm for stock prediction modeling. *Neurocomputing* **142** (October 2014) 48–59
14. Adraghi, K.P., Cook, R.D.: Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **367**(1906) (November 2009) 4385–4405
15. Granger, C.W.J.: Testing for causality. *Journal of Economic Dynamics and Control* **2** (January 1980) 329–352
16. Schreiber, T.: Measuring Information Transfer. *Physical Review Letters* **85**(2) (July 2000) 461–464
17. Barnett, L., Barrett, A.B., Seth, A.K.: Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical Review Letters* **103**(23) (December 2009)
18. Piotr, P., Youssef, H., Alain, C., Lakhil, L.: Improving multivariate time series forecasting with random walks with restarts on causality graphs. In: *ICDM Workshops 2017*.