GP-Based Electricity Price Forecasting

Alberto Bartoli, Giorgio Davanzo, Andrea De Lorenzo, and Eric Medvet

DIII, University of Trieste, Via Valerio, Trieste, Italy

Abstract. The electric power market is increasingly relying on competitive mechanisms taking the form of day-ahead auctions, in which buyers and sellers submit their bids in terms of prices and quantities for each hour of the next day. Methods for electricity price forecasting suitable for these contexts are crucial to the success of any bidding strategy. Such methods have thus become very important in practice, due to the economic relevance of electric power auctions.

In this work we propose a novel forecasting method based on Genetic Programming. Key feature of our proposal is the handling of outliers, i.e., regions of the input space rarely seen during the learning. Since a predictor generated with Genetic Programming can hardly provide acceptable performance in these regions, we use a classifier that attempts to determine whether the system is shifting toward a difficult-to-learn region. In those cases, we replace the prediction made by Genetic Programming by a constant value determined during learning and tailored to the specific subregion expected.

We evaluate the performance of our proposal against a challenging baseline representative of the state-of-the-art. The baseline analyzes a real-world dataset by means of a number of different methods, each calibrated separately for each hour of the day and recalibrated every day on a progressively growing learning set. Our proposal exhibits smaller prediction error, even though we construct one single model, valid for each hour of the day and used unmodified across the entire testing set. We believe that our results are highly promising and may open a broad range of novel solutions.

1 Introduction

The electric power industry has shifted from a centralized structure to a distributed and competitive one. In many countries of the world, electricity markets of several forms have been established that allow consumers to select among different providers according to reliability and cost metrics. Although the legal and technical features of such markets are regulated differently in each country, the presence of *auctions* in which buyers and sellers submit their bids in terms of prices and quantities is commonplace [13].

An important form of such auctions can be found in the *day-ahead* market, in which producers and consumers present price-sensitive supply offers and demands for each hour of the next day. Each day a coordinating authority determines the outcome of the auction in terms of electricity flows and final prices.

S. Silva et al. (Eds.): EuroGP 2011, LNCS 6621, pp. 37–48, 2011.

The economic relevance of these auctions makes the ability to accurately predict next-day electricity prices very important in practice, for both producers and consumers: bidding strategies are based on price forecast information hence the actual benefit obviously depends heavily on the accuracy of such information. Not surprisingly, thus, many approaches to electricity price forecasting have been explored in the recent years [15,3,2,1,11,10,8,4].

In this work we examine the usefulness of Genetic Programming (GP) in the context of day-ahead electricity price forecasting. We propose two GP approaches that differ in the choice of the variables used as input of the evolutionary search and two hybrid approaches. The hybrid approaches augment the GP-built predictor with a classifier that predicts the interval to which the next prices will belong. When the predicted interval was rare in the learning set, the output of the GP-based predictor is replaced by the mean value that was observed, in the learning set, for the predicted interval. The rationale for this design is that the GP-generated predictor can hardly provide acceptable performance in regions of the input space that have been rarely seen during the learning. The classifier attempts to determine whether the system is shifting toward the difficult-to-learn region, in which case we simply predict a constant value tailored to the specific subregion expected.

We assess our results by comparing them to a very challenging baseline that, in our opinion, may be considered as representative of the state-of-the-art for the problem. The dataset consists of hourly market clearing prices set by the California Power Exchange (CalPX) from July 5, 1999 to June 11, 2000. This period includes an important market crisis, which started on May 1, 2000, that provoked significant volatility and large variations in prices, due to bankruptcy and strong financial problems of major players in the market [9]. The forecasting methods used as baseline are those discussed in [14], which evaluates the performance of 12 widely different approaches proposed in the literature. For each approach, 24 different models are constructed and carefully tuned, one for each hour of the next day. Each model is recalibrated every day, by shifting the end of the learning set to the current day—thereby growing the learning set every day. We also include in the comparison the results from [9], which apply 4 AI-based methods to the same dataset. Even in this case, each method is calibrated differently for each hour of the day and is recalibrated every day.

We evaluate the performance of our predictors with the same error index used in these works and obtain very interesting results. The GP-based approaches exhibit slightly worse performance than those of the traditional methods. The hybrid approaches, on the other hand, provide *better* performance. In fact, they even provide a performance better than a conceptual (not implementable) forecasting method obtained by selecting in each week of the testing set the best of all the other predictors for that week.

We remark that our results have been obtained in a scenario more challenging than the baseline: (i) we construct one single predictor, valid for every hour of each day; and (ii) we never recalibrate our predictor, i.e., we use the very same learning set used in [14,9] at the beginning of the simulation and then we leave the predictor unmodified across the entire testing set.

We believe our contribution is relevant for several reasons. First, we provide a novel solution to a problem highly relevant in practice that compares favorably to the current state-of-the-art. Second, we extend the set of application domains in which GP may outperform, or at least compete with, traditional approaches. Third, we show a simple yet effective way to cope with a dataset that do not cover the output space uniformly.

2 Our Approach

2.1 Overview

We propose two techniques for day-ahead electricity price forecasting. The first technique is entirely GP-based (Section 2.2), while the second one is a hybrid technique that combines the output of the GP-generated predictor with the output of a second simple predictor, to be used when the system is shifting toward regions that have been rarely seen during learning (Section 2.3).

We denote by P_h the observed price for hour h and by $\hat{P}_h$ the predicted price for that hour.

Every day at midnight the prediction machinery generates a forecast $\hat{P}_h$ for the following 24 hours, i.e., for each $h \in \{1, \ldots, 24\}$. This is the usual pattern used in the literature, although in practice prediction occurs around mid-day, not at midnight.

The variables potentially available for generating $\hat{P}_h$ are:

- $P_{h-24}, \ldots, P_{h-168}$, that represent the previously observed values for the price (e.g., P_{h-168} indicates the observed price one week before the generation of the prediction).
- $H_{h-24}, H_{h-48}, H_{h-72}, H_{h-96}, H_{h-120}, H_{h-144}$ and H_{h-168}, that represent the maximum value observed for the price in the corresponding day (e.g., H_{h-48} indicates the maximum price in the day that precedes the generation of the prediction).
- $I_{h-24}, I_{h-48}, I_{h-72}, I_{h-96}, I_{h-120}, I_{h-144}$ and I_{h-168}, that represent the minimum value observed for the price in the corresponding day.
- $N_h, N_{h-1}, \ldots, N_{h-168}$, a set of binary values that represent whether an hour corresponds to night-time, i.e., $N_k = 1$ if $1 \leq k \leq 5$ and $N_k = 0$ otherwise.
- $I_h, I_{h-1}, \ldots, I_{h-168}$, a set of binary values that represent whether an hour corresponds to holidays.
- An enumerated variable $h \in \{1, 2, \ldots, 24\}$ that represents the hour of the day for $\hat{P}_h$.
- An enumerated variable $d \in \{1, 2, \ldots, 7\}$ that represents the day of the week for $\hat{P}_h$ (from Sunday to Saturday).

We remark that we rely only on measured values for the variable to be predicted, i.e., we do not require any exogenous variable. Existing literature, in contrast, often assumes that the prediction machinery has some exogenous variables available, e.g., temperature, actual or forecasted load and alike. Indeed, 6

of the 12 models in [14] use load forecast as exogenous variable. We believe that our approach may be more practical, simpler to implement and less dependent on "magic" tuning numbers—e.g., if temperature were to be used, where and at which hour of the day itshould be taken?

We partition the dataset in three consecutive time intervals, as follows: the *training set*, for performing the GP-based evolutionary search; the *validation set*, for selecting the best solution amongst those found by GP; the *testing set*, for assessing the performance of the generated solution.

2.2 GP Approach

The set of variables potentially available to the prediction system is clearly too large to be handled by the GP search efficiently. We consider two configurations: one, that we call *GP-baseline*, in which the terminal set consists of the same variables used in the best-performing method of the baseline work (except for any exogenous variable, such as the load) [14]. The resulting terminal set is: $\{P_{h-24}, P_{h-48}, P_{h-168}, I_h, N_h, L_{h-24}\}$. The other configuration, that we call *GP-mutualInfo*, uses a terminal set that consists of variables selected by a feature selection procedure that we describe below.

The procedure is based on the notion of *mutual information* between pairs of random variables, which is a measure of how much knowing one of these variables reduces the uncertainty about the other [12]. The procedure consists of an iterative algorithm based on the training and validation portions of the dataset, as follows. Set S initially contains all the 498 variables potentially available to the prediction system. Set S_{out} is initially empty and contains the selected variables to be used for the GP search.

1. Compute the mutual information m_i between each variable $X_i \in S$ and the price variable Y.
2. For each pair of variables $X_i \in S, X_j \in S$, compute their mutual information m_{ij}.
3. Let $X_i \in S$ be the variable with highest m_i. Assign $S := S - X_i$ and $S_{out} := S_{out} + X_i$. For each variable $X_j \in S$, modify the corresponding mutual information m_j as $m_j := m_j - m_{ij}$.
4. Repeat the previous step until S_{out} contains a predefined number of elements.

We chose to execute this feature selection procedure for selecting 8 variables. The resulting terminal set to be submitted to GP is:

$$\{P_{h-24}, P_{h-168}, I_h, I_{h-24}, I_{h-168}, N_h, H_{h-24}, L_{h-24}, h, d\}$$

At this point we run the GP search on the training set, with parameters set as described in Section 3.2. Next, we compute the fitness of all individuals on the validation set. Finally, we select the individual that exhibits best performance on this set as predictor. This individual will be used along the entire testing set.

2.3 Hybrid Approach

Our hybrid approach generates a GP-based predictor exactly as described in the previous section, but introduces an additional component to be used in the testing phase. This component is synthesized using the training and validation portions of the dataset, as follows.

1. We define 10 equally-sized intervals for the observed price values in the training and validation set and define each such interval to be a class.
2. We compute the mean value for each class.
3. We execute a feature selection procedure [6] consisting of a genetic search algorithm [5] and select 95 of the 498 variables potentially available.
4. We train a classifier for the above classes based on the variables selected at the previous step. In other words, this classifier predicts the class to which the next price value will belong. In our experiments we have used a multilayer perceptron.

The choice of the specific algorithms used at step 3 and 4 has been influenced by the software tool used for this purpose (Weka [7]).

In the testing phase, the prediction is generated as follows. We denote by C_A the set of the 2 classes with more elements and by C_B the set of the other classes. Let $\hat{c}$ be the predicted class for P_i. If $\hat{c} \in C_A$ then the predicted value $\hat{P}_i$ is the value generated by the GP predictor, otherwise $\hat{P}_i$ is the mean value computed for $\hat{c}$ (step 2 above).

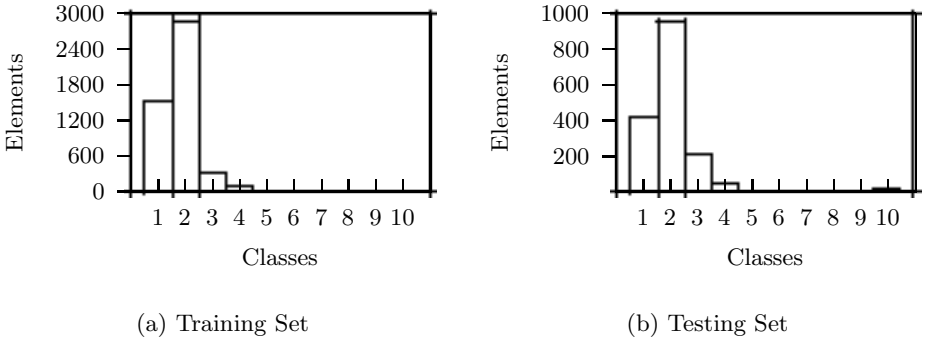
The rationale of this design is that the GP-generated predictor cannot be expected to perform well in regions of the input space that have been rarely seen during the learning. The classifier attempts to determine whether the system is shifting toward the difficult-to-learn region, in which case we simply predict a constant value determined during training and tailored to the specific subregion expected.

Figures 1(a) and 1(b) show the distributions of price values in the training set and in the testing set, respectively (in the validation set all values happen to belong to the first 2 classes). The percentage of elements in the 2 classes with more elements is 92% in the learning set (training and validation) and 82% in the testing set.

3 Experimental Evaluation

3.1 Dataset and Baseline

As clarified in the introduction, we believe the dataset and baseline that we have used are highly challenging and may be considered as representative of the state-of-the-art. The dataset consists of hourly market clearing prices set by the California Power Exchange (CalPX) from July 5, 1999 to June 11, 2000 (Figure 2). This period includes a market crisis period characterized by large price volatility, that started on May 1, 2000 and lasted beyond our dataset [9].



(a) Training Set (b) Testing Set

Fig. 1. Distribution of price values in classes

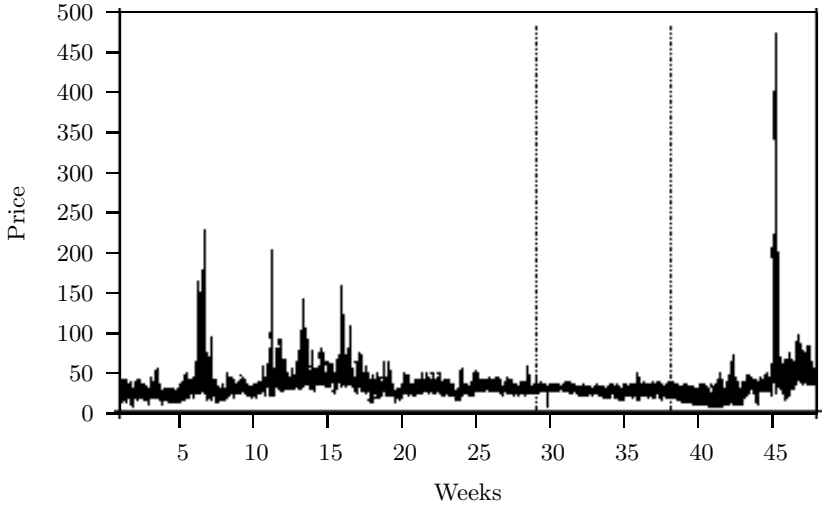


Fig. 2. Dataset used for evaluating the proposed methods. The vertical line at the right indicates the division between learning set and testing set. The vertical line at the left indicates the division between training set and validation set (used only in our approaches, see Section 3.2).

We use the results from [14] as main baseline. This work examines a set of widely differing approaches proposed earlier in the literature: basic autoregressive (AR), spike preprocessed (p-AR), regime switching (TAR), mean-reverting jump diffusion (MRJD), iterated Hsien-Manski estimator (IHMAR), smoothed non-parametric maximum likelihood (SNAR) (please refer to the cited work for full details). Each approach is applied with and without load forecast as exogenous variable. For each approach, 24 different models are constructed and carefully tuned, one for each hour of the next day. In the testing phase each model is recalibrated every day, by shifting the end of the learning set to the current day—thereby growing the learning set every day. The initial learning set contains the first 9 months, from July 5, 1999, to April 2, 2000. The next 10 weeks constitute the testing set, which thus includes the market crisis mentioned above[1].

We include in the comparison also the results from [9], which applies several AI-based approaches to the same dataset[2]: autoregressive neural network (ANN), local regression (LOCAL), linear regression tree (TREE), generalized additive model (GAM) (again, please refer to the cited work for full details). This work follows the same structuring as the previous one: it uses the same learning set, it models each hour of the day separately and recalibrates each model every day.

The performance index is the Weekly-weighted Mean Absolute Error (WMAE), defined as:

$$\text{WMAE} = \frac{\sum_{h=1}^{168} |P_h - \hat{P}_h|}{\sum_{h=1}^{168} P_h}$$

where P_h is the actual price for h and $\hat{P}_h$ is the predicted price for that hour.

3.2 Settings

We split the dataset as in [14,9]: the learning set contains the first 9 months, whereas the next 10 weeks constitute the *testing set*. We further split the learning data in two consecutive intervals used as described in Section 2.2: a *training set* from July 5, 1999, to January 30, 2000, is used for the GP search; a *validation set* from January 31, 2000 to April 2, 2000, is used for selecting the best individual produced by the GP search.

We used WMAE measured on the training set as fitness function. We could have used other fitness functions (e.g., squared error distances) and then assess the resulting performance on the testing set based on the performance index of interest in this work, i.e., WMAE. We have not explored this possibility in depth, but preliminary results suggest that there are no substantial differences between these two approaches. Indeed, this finding is in line with a similar assessment made in [14].

[1] The cited work analyzes also another dataset from the Nordic Power Exchange (http://www.nordpoolspot.com/) augmented with hourly temperatures in Sweden. We have not yet applied our approach to this dataset.

[2] This work actually considers a longer testing set. We include here the results for the same testing set used in [14].

We experimented with four configurations: GP-baseline, GP-mutualInfo, Hybrid-baseline (i.e., coupled with GP-baseline), Hybrid-mutualInfo. The GP searches have been made with the same set of parameters, except for the composition of the terminal set, that is different for the cases GP/Hybrid-baseline and GP/Hybrid-mutualInfo (see Section 2.2). The functions set includes only the four basic arithmetic operators and the terminal set always includes a few basic constants: 0.1, 1, 10. During our early tests we experimented with different combinations of population size and number of generations. Concerning the former, we swept the range between 500 and 1000 individuals, keeping fixed the number of generations, and found no significant differences in WMAE performance. However, we also found that a population with 1000 individuals triplicates the computation time required by one with 500 individuals, thus we decided to select 500 as population size. Concerning the number of generations, we decided to use 1200 generations after some exploratory experimentation. The full set of GP-related parameters is summarized in Table 1.

Table 1. GP parameters

Parameter	Settings
Populations size	500
Selection	Tournament of size 7
Initialization method	Ramped half-and-half
Initialization depths	1
Maximum depth	5
Elitism	1
Reproduction rate	5%
Crossover rate	80%
Mutation rate	15%
Number of generations	1200

For each configuration (i.e., GP-baseline, GP-mutualInfo, Hybrid-baseline, Hybrid-mutualInf): (i) we ran 128 GP executions, each with the parameters in Table 1; (ii) at the end of each execution we selected the individual with the best fitness on the training set, thereby obtaining a final population of 128 individuals; (iii) we evaluated the fitness of these individuals on the validation set and selected the one with best fitness as predictor to be used in the testing set. Concerning the hybrid approach, we used the Weka tool in the standard configuration [7] and experimented with several forms of classifier: Random Tree, Random Forest Tree, SVM, Multilayer Perceptron. The latter is the one that exhibited best performance and has been used for deriving the results presented here.

Finally, a few notes about execution time: each GP search took about 34 hours on 4 identical machines running in parallel, each machine being a quad-core Intel Xeon X3323 (2.53 GHz) with 2GB RAM; the training of the classifier took about 1 hour on a single core notebook (2 GHz), with 2GB RAM; the variable selection procedure (Section 2.2) took a few minutes on the same notebook.

3.3 Results

Table 2 presents the salient results. The first four rows contain the average WMAE along the testing set for each of the approaches that we have developed. To place these results in perspective, the next set of rows provides the same index, extracted from [14]. In particular, the first 12 rows correspond to the 6 approaches, each tested with and without predicted load as exogenous variable (models with the exogenous variable are denoted by the X suffix). Then, we provide the mean for the full set of 12 models, the mean for the 6 pure-price models only and the mean for the 6 models with exogenous variable. Finally, the row labeled Ideal gives the mean WMAE of an optimal (merely conceptual) model, constructed by selecting the best performing model in each week of the testing phase (the cited work provides the WMAE of each model in each week). The final set of rows provides the corresponding WMAE values from [9]. We excluded method LOCAL from the evaluation of mean values, as it is clearly an outlier. The row labeled Ideal has the same meaning as above whereas the row IdealBoth corresponds to selecting the best model in each week from the full set of 16 predictors provided by the cited works.

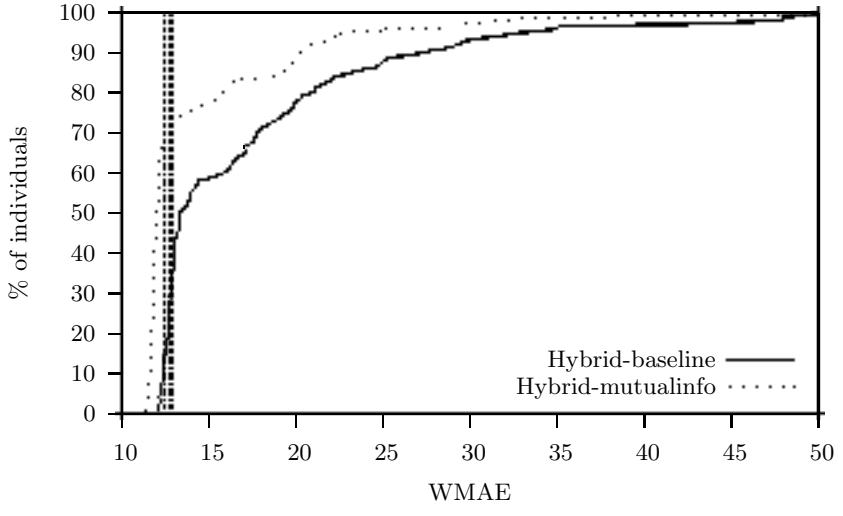
The key result is that both the hybrid methods perform better than all the other methods, including the "optimal" (and merely conceptual) predictors constructed by selecting the best predictor in each week. We believe this is a very promising result. The fact that our approaches construct one single model valid for every hour of the day and that we never recalibrate our models along the entire testing set, may only corroborate this claim.

Table 2. Mean WMAE results in the testing set. The upper portion of the left table corresponds to our approaches, the lower portion are results from [9] (the huge value for the LOCAL method is not a typing mistake), the right table are results from [14].

Method	Mean WMAE (%)
GP-mutualInfo	20.70
GP-baseline	16.17
Classifier-base	16.03
Hybrid-mutualInfo	11.84
Hybrid-baseline	12.32
ANN	13.11
LOCAL	154499.01
TREE	14.02
GAM	13.29
Mean	13.47
Ideal	12.83
Ideal both	12.42

Method	Mean WMAE (%)
AR	13.96
ARX	13.36
p-AR	13.44
p-ARX	12.96
TAR	13.99
TARX	13.31
MRJD	15.39
MRJDX	14.67
IHMAR	14.01
IHMARX	13.37
SNAR	13.87
SNARX	13.17
Mean	13.79
Mean pure-price only	14.11
Mean with load only	13.47
Ideal	12.64

Fig. 3. Distribution of mean WMAE performance for the final populations, with hybrid methods. Vertical lines indicate the WMAE for the three Ideal methods shown in Table 2.



In order to gain further insights into the ability of our hybrid methods to effectively generate accurate predictors, we evaluated WMAE across the entire testing set for all the individuals of the final population. That is, rather than selecting one single individual based on its WMAE performance in the validation set, we take all the individuals. The distribution of the respective WMAE performance is shown in Figure 3. Vertical lines indicate the WMAE for the three Ideal methods shown in Table 2.

It can be seen that the better performance exhibited by the hybrid methods is not an occasional result provoked by a single "lucky" individual: these methods consistently tend to generate individuals whose performance is better than any of the baseline methods. Indeed, the baseline performance is improved by more than half of the final population (Hybrid-baseline) and by approximately three-quarters of the final population (Hybrid-mutualInfo).

For completeness of analysis, we assessed the impact of the classifier from several points of view. Concerning the prediction mistakes performed by the classifier in the testing set, it provoked a wrong replacement of the prediction by GP 2.5% of the times, and it provoked a wrong use of the prediction by GP 10.11% of the times. The performance (mean WMAE in the testing set) that one could obtain by our Hybrid approach implemented by a perfect classifier—i.e., one that never makes any prediction mistake in the testing set—is 10.50% (mutualInfo) and 10.96% (baseline). Finally, the performance that one could obtain by always using the mean value for the class predicted by a perfect classifier is 15.49%.

From these data we observe what follows. First, attempting to improve the prediction accuracy of the classifier further is probably not worthwhile (a perfect

classifier does not deliver a substantial improvement over Hybrid-mutualInfo). Second, our hybrid approach indeed boosts performance of its building blocks—classifier-based prediction and GP-based prediction: the former is slightly better than the latter, but their combination is substantially better than either of them. Third, the simple classifier-based prediction exhibits performance that is better than GP-only methods and is only slightly worse than the 16 baseline methods.

4 Concluding Remarks

We have proposed novel GP-based methods for electricity price forecasting that are suitable for day-ahead auctions. We designed simple yet effective mechanisms for enabling GP to cope with the strong volatility that may be typical of this difficult application domain.

We assessed our proposal on a challenging real-world dataset including a period of market crisis, and compared our results against a baseline that is representative of the current state-of-the-art. Our hybrid methods performed better than all the 16 methods considered, and better than ideal (not implementable) predictors constructed by taking the best of those predictors in each week. We also showed that our methods tend to systematically generate predictors with good performance: we actually generated tens of predictors that exhibit better performance than those used as baseline.

Although our approach has certainly to be investigated further, in particular on other datasets, we believe that our results are significant and highly promising.

Acknowledgments

We are very grateful to Cyril Fillon for having initiated our group into the secrets (and power) of GP, for the development of Evolutionary Design (the API used in this research) and for his comments on this work. We are also grateful to Gabriele Del Prete for his hard work during the initial stage of this research.

References

1. Amjady, N., Keynia, F.: Day ahead price forecasting of electricity markets by a mixed data model and hybrid forecast method. International Journal of Electrical Power and Energy Systems 30(9), 533–546 (2008)
2. Areekul, P., Senjyu, T., Toyama, H., Yona, A.: A hybrid ARIMA and neural network model for Short-Term price forecasting in deregulated market. IEEE Transactions on Power Systems 25(1), 524–530 (2010)
3. Catalao, J.P.S., Pousinho, H.M.I., Mendes, V.M.F.: Hybrid Wavelet-PSO-ANFIS approach for Short-Term electricity prices forecasting. IEEE Transactions on Power Systems (99), 1–8 (2010)
4. Cuaresma, J.C., Hlouskova, J., Kossmeier, S., Obersteiner, M.: Forecasting electricity spot-prices using linear univariate time-series models. Applied Energy 77(1), 87–106 (2004)

5. Goldberg, D.E.: Genetic algorithms in search, optimization and machine learning (1989)
6. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: Proc. 17th Intern. Conf. Machine Learning, pp. 359–366 (2000)
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations 11(1) (2009)
8. Koopman, S.J., Ooms, M.: Forecasting daily time series using periodic unobserved components time series models. Computational Statistics & Data Analysis 51(2), 885–903 (2006)
9. Mendes, E.F., Oxley, L., Reale, M.: Some new approaches to forecasting the price of electricity: a study of californian market, `http://ir.canterbury.ac.nz/handle/10092/2069`, RePEc Working Paper Series: No. 05/2008
10. Mount, T.D., Ning, Y., Cai, X.: Predicting price spikes in electricity markets using a regime-switching model with time-varying parameters. Energy Economics 28(1), 62–80 (2006)
11. Pedregal, D.J., Trapero, J.R.: Electricity prices forecasting by automatic dynamic harmonic regression models. Energy Conversion and Management 48(5), 1710–1719 (2007)
12. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
13. Sheblé, G.B.: Computational auction mechanisms for restructured power industry operation. Springer, Netherlands (1999)
14. Weron, R.: Misiorek: Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. International Journal of Forecasting, 744–763 (2008)
15. Wu, L., Shahidehpour, M.: A hybrid model for Day-Ahead price forecasting. IEEE Transactions on Power Systems 25(3), 1519–1530 (2010)