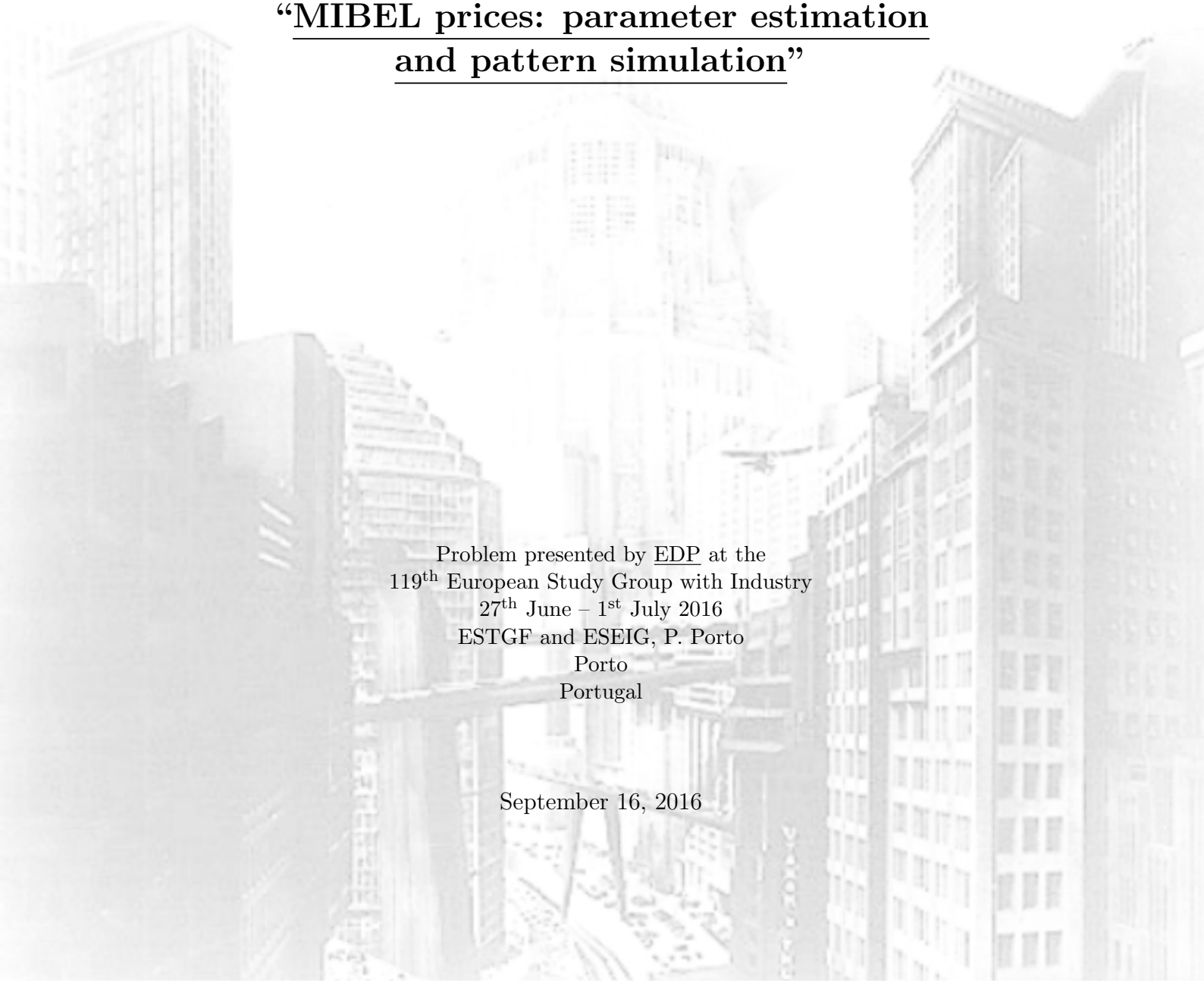
Portuguese Study Groups' Reports

Report on "MIBEL prices: parameter estimation and pattern simulation"



Problem presented by EDP at the
119th European Study Group with Industry
27th June – 1st July 2016
ESTGF and ESEIG, P. Porto
Porto
Portugal

September 16, 2016

Problem presented by: Ricardo Covas (EDP)

Study group contributors: Ana Borges, Andreia Monteiro, Eliana Costa e Silva, Marina Andrade, M. Filomena Teodoro, and Raquel Menezes

Report prepared by: Ana Borges (aib@estg.ipp.pt)
Eliana Costa e Silva (eos@estg.ipp.pt)
Marina A. P. Andrade (marina.andrade@iscte.pt),
and
M. Filomena Teodoro (maria.alves.teodoro@marinha.pt)

Executive summary

EDP Group is an Energy Solutions Operator which operates in the business areas of generation, supply and distribution of electricity and supply and distribution of gas.

The challenge proposed by EDP consists in simulating electricity prices not only for risk measures purposes but also for scenario analysis in terms of pricing and strategy. Data concerning hourly electricity prices from 2008 to 2016 was provided.

Numerous methods to deal with Electricity Price Forecasting (EPF) have been proposed and can be classified as: (i) multi-agent models, (ii) fundamental models, (iii) reduced-form models, (iv) statistical models and (v) computational intelligence models. A recent an exaustive review is presentes in [13].

During this study group different promising Statistical techniques were propose by the study group contributors: ARIMA, sARIMA, Longitudinal Models, Generalized Linear Models and Vector Autoregressive Models. In this report a GLM and a vector autoregressive model are presented and their predictive power is discussed.

In the GLM framework two different transformations were consider and for both the season of the year, month or winter/summer period revealed significant explanatory variables in the different estimated models.

On the other hand, the multivariate approach using VAR considering as exogenous variables the meteorologic season and the type of day yield a multivariate model that explains the intra-day and intra-hour dynamics of the hourly prices. Although the forecast do not exactly replicate the real price they are quite similar.

In both of the approaches here reported a more extensive work would certainly improve the proposed models.

In conclusions, EPF is a growing area that groups multiple different approaches that can be applied. In fact, other approaches from multi-agent models, fundamental models, reduced-form models and computational intelligence models, also present a great space for EPF.

1 Introduction

Under the 109th ESGI, EDP - Energias de Portugal submitted a mathematical challenge titled *MIBEL prices - parameter estimation and pattern simulation*. EDP Group is an Energy Solutions Operator which operates in the business areas of generation, supply and distribution of electricity and supply and distribution of gas. EDP with nearly 14 000 MW (2012 update and excluding wind power) of installed capacity in the MIBEL[1] (Iberian Electricity Market), is the only company in the Iberian Peninsula with generation, distribution and supply (both electricity and gas) activities in Portugal and Spain.

The challenge proposed consists in simulating electricity prices not only for risk measures purposes but also for scenario analysis in terms of pricing and strategy.

There are numerous methods proposed to deal with Electricity Price Forecasting (EPF). Weron in [13] explains in his exhaustive review article the complexity of available methods, revealing their strengths and weaknesses, reducing them into five major categories: [(i)] multi-agent models, [(ii)] fundamental models, [(iii)] reduced-form models, [(iv)] statistical models and [(v)] computational intelligence models.

During the 119th ESGI the study group contributors focused on statistical approaches. Most of the statistical approaches consist in methods that forecast the current electricity price by using a mathematical combination of the previous prices and/or previous or current values of exogenous factors, such as, consumption and production figures, or weather variables (see [13] for further detail).

Statistical EPF models are mainly inspired from economics literature such as game theory models and time-series econometric models, as explained also by [9], where they present an extremely relevant a summary of selected finance and econometrics inspired literature on spot electricity price forecasting (see Table 3 in [9]).

Synthesizing, in EPF, and confining this study mainly to short term price forecasting, autoregression models are widely used such as, the univariated AutoRegressive Average model (ARMA), a standard time series model that takes into account the random nature and time correlations of the phenomenon under study, the AutoRegressive Integrated Moving Average (ARIMA), an extension of ARMA that enables a transformation of the series to the stationary form, or even the seasonal ARIMA model (SARIMA), that captures a possible existence of seasonality. The forecasting of ARMA-type models can be conducted via the Durbin-Levinson algorithm or the innovations algorithm, or by using the Kalman filter for models specified in state space form. As electricity prices can be influenced by the present and

[1] http://www.mibel.com/

past values of various exogenous factors, such as generation capacity, load profiles and ambient weather conditions [13], usually it is used and extension of the previous mentioned time series models with exogenous or input variables such as ARX, ARMAX, ARIMAX and SARIMAX (see e.g. [12] for a excellent manual in Time Series Analysis). This approach was also consider by the study group contributors but not addressed due to time limitations. This is an approach that deserves further attention.

EPF literature has mainly concerned on models that use information at daily level, however this particularly problem proposed is interested in forecasting intra-day prices using hourly data (disaggregated data), and therefore, it is necessary to consider models that explore the complex dependence structure of the multivariate price series. For that, a vector autoregressive structure (VAR) approach has been recently proposed and it will be detailed latter and applied to the data in study in Section 4.

Multivariate time series analysis is used when one wants to model and explain the interactions and co-movements among a group of time series variables, such as, Consumption and income; Stock prices and dividends; Forward and spot exchange rates; Interest rates, money growth, income, inflation. In this scope [3], [11], [4] have proposed some techniques: VAR, MAR, Auto Regressive Moving Average Vector (VARMA), GARCH, ARFNN- fusion of VAR and fuzzy neural networks (caotic, outliers), Extended Kalman Filter, Polynomial fitting.

Temporal Distribution Extrapolation is another possible approach. It considers the kernel density estimation taking into account, for example, pseudopoints. It is a nonparametric technique which estimates the distribution of a random (univariate ou multivariate) variable minimizing some measure (see figure 1). Quite interesting work is presented in [5], [6].

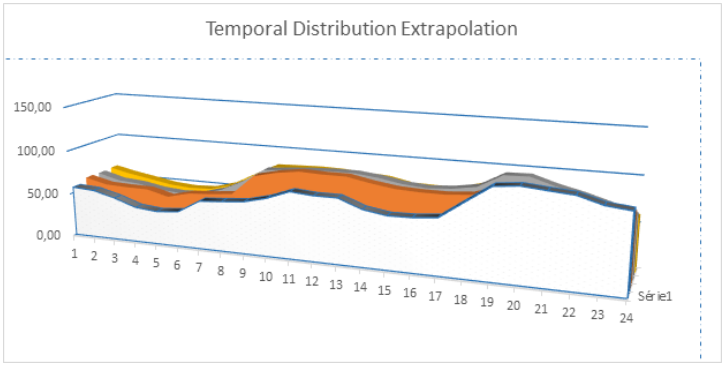


Figure 1: Temporal distribution extrapolation

Another approach proposed by this study group was Longitudinal/panel models, in particular linear mixed effect models as described in [2]. Longitu-

dinal data is usually characterized as response variables that are measured repeatedly through time for a group of individuals, in this case, a group of hours. The main characteristic of longitudinal models is that they model both the dependence among the response on the explanatory variables and the autocorrelation among the responses. If we, in fact, consider the progression of the price of each hour independent or, at least, the independence of groups of hours (grouping hours accordingly, for example, to dual hourly rate), this type of models are powerful tool to model the progression of the price through time.

In the rest of this section more details on the challenge proposed by EDP and on the data provided will be presented. Section 2 starts by presenting the exploratory analysis of the datasets provided by EDP and continues with the study on the co-variables that may predict the hourly prices pattern. In Section 3 is presented a Generalized Linear model approach. Next on Section 4 a multivariate approach to the forecast of disaggregated electricity prices is presented and the daily patterns of the predicted values are compared against the real observed values. Finally in Section 5 conclusions are drawn and suggestions for future work are pointed.

1.1 Challenge and aims

The challenge proposed by EDP was:

> The daily market electricity prices,
>
> $$Y_t = [y_{1t}, y_{2t}, \ldots, y_{nt}]$$
>
> is a strip of prices (one for each hour of the day), all simultaneously observed once at a given time of each day.
>
> Therefore the daily market prices can be interpreted as a time dependent multivariate random variable. For simplification, we may suppose that $Y_t, \quad t = 1, 2, \ldots,$ is multivariate normal random variable with an inner variance-covariance matrix (constant in time) and with an auto-regressive structure for time dependence.
>
> Although simulating multivariate normal distributions is a straightforward exercise we need (and purpose) to simulate it subject to restrictions on the sum of $Y_t, \quad t = 1, 2, \ldots.$
>
> We also welcome some insights on the estimation side when this conditioning on the sum of $Y_t, \quad t = 1, 2, \ldots$ is made.

An example of the hourly prices is presented in Figure 1.1 for the 31 days in January 2016. This figure appears do display a common pattern among the different hours of the day.

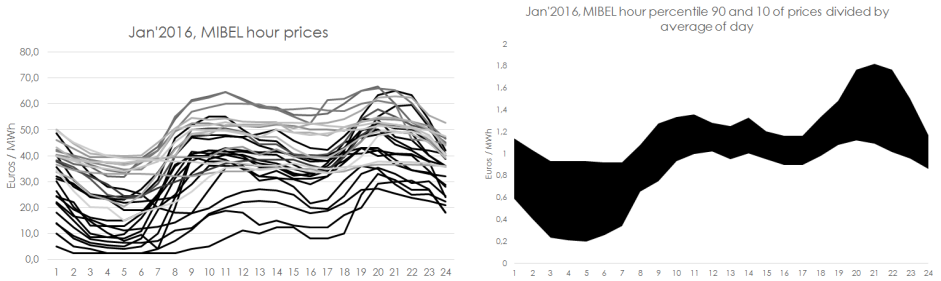


Figure 2: On the *left*, MIBEL prices from January 2016. On the *right*, 10^{th} up to 90^{th} percentiles MIBEL hourly prices divided by price average of each day MIBEL hour from 10^{th} up to 90^{th} percentiles of prices divided by price average of day (see also Figure 3). (provided by EDP during 119th ESGI).

1.2 Database

EDP provided data from electricity price from January 2008 to June 2016. At the beginning of the week the data concerned the ratios of the hourly prices by the average price of each day, in a total 3102 observations of the 24 hours of the day (23 or 25). Figure 3 presents the first observations.
A second dataset, with disaggregated data, i.e., hourly prices and average day price, was provided on June, 28th (see Figure 1.2).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Data	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	01/01/2008	1.52	1.53	1.48	1.34	1.09	0.85	0.84	0.82	0.8	0.71	0.72	0.72	0.82	0.82	0.84	0.82	0.82	0.82	0.95	1.09	1.12	1.18	1.09	1.2	
3	02/01/2008	0.92	0.83	0.8	0.71	0.59	0.61	0.82	0.83	0.92	1.19	1.33	1.33	1.3	1.24	1.07	0.95	0.95	1.02	1.32	1.26	1.17	1.1	0.89	0.87	
4	03/01/2008	0.83	0.8	0.71	0.59	0.55	0.59	0.8	0.83	0.86	0.96	1.1	1.07	1.07	0.93	0.87	0.88	0.91	1.19	1.46	1.5	1.46	1.43	1.32	1.28	
5	04/01/2008	0.82	0.74	0.73	0.73	0.64	0.73	0.75	0.78	1.08	1.16	1.21	1.19	1.19	1.13	1.08	1.06	1.06	1.09	1.21	1.21	1.18	1.13	1.06	1.04	
6	05/01/2008	1.02	1.1	0.88	0.81	0.74	0.74	0.73	0.75	0.81	0.9	1.17	1.17	1.17	1.1	0.95	0.88	0.86	0.89	1.31	1.31	1.2	1.19	1.13	1.17	
7	06/01/2008	1.02	1.01	0.82	0.76	0.62	0.62	0.61	0.56	0.62	0.72	0.88	0.92	0.99	0.99	0.95	0.89	0.85	0.99	1.41	1.46	1.63	1.63	1.53	1.53	
8	07/01/2008	1.06	0.74	0.7	0.65	0.63	0.64	0.74	0.78	1.02	0.95	1.08	1.1	1.11	1.09	1.07	1.06	1.02	1.08	1.28	1.3	1.29	1.29	1.18	1.15	
9	08/01/2008	0.87	0.76	0.73	0.67	0.63	0.64	0.73	0.91	1.13	1.16	1.15	1.18	1.15	1.1	1.06	1.06	1.06	1.15	1.32	1.29	1.18	1.18	0.99	0.91	
10	09/01/2008	0.82	0.74	0.71	0.65	0.62	0.64	0.74	0.9	1.11	1.13	1.16	1.17	1.17	1.14	1.07	1.05	1.07	1.13	1.33	1.34	1.23	1.14	1.04	0.92	
11	10/01/2008	0.9	0.82	0.76	0.7	0.67	0.68	0.79	0.98	1.16	1.17	1.18	1.17	1.15	1.12	1.05	1.04	1.05	1.15	1.23	1.23	1.17	1.1	0.9	0.86	
12	11/01/2008	0.82	0.8	0.7	0.59	0.57	0.57	0.78	0.91	1.2	1.21	1.24	1.21	1.17	1.12	0.93	0.98	0.99	1.13	1.24	1.28	1.22	1.16	1.13	1.05	
13	12/01/2008	1	0.92	0.86	0.83	0.78	0.78	0.78	0.85	0.87	0.95	1.12	1.11	1.12	1.02	0.94	0.91	0.9	0.96	1.24	1.25	1.27	1.24	1.08	1.23	

Figure 3: Ratio of hourly prices by the average daily price.

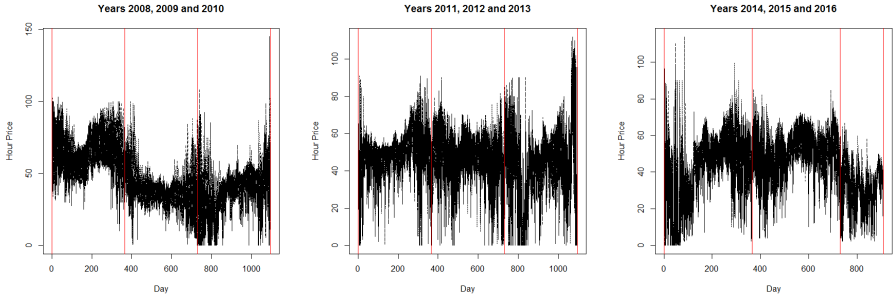


Figure 4: Hourly prices for the period from January 2008 to June 2016.

During the first discussions, the following question came up: "Can we reduce the number of components of Y_t?". I.e., are there significant differences between Y_i and Y_j, for $i \neq j$?
To answer to this question on both of the datasets the following candidates to co-variables were added:

- Day of the week – $C_1 = 0, 1, 2, 3, 4, 5, 6$ (Mon, ..., Sunday)
- Weekday/Saturday/Sunday – $C_2 = 0, 1, 2$
- Weekday/Weekend – $C_3 = 0, 1$
- Regular day/ holiday – $C_4 = 0, 1$
- Season – $C_5 = 0, 1, 2, 3$ (Winter, Spring, Summer, Autumn)
- Month – $C_6 = 0, \ldots, 11$ (Jan, ..., Dec)
- Summer/Winter Hour – $C_7 = 0, 1$

The importance of such covariables is analyzed in the next sections.

2 Exploratory Analysis

In an initial exploratory analysis, the data originally provided (in what follows named rescaled data) revealed serious problems which can be confirmed in some of next figures. Such problems induced us to contact again the company explaining our worries. Consequently, we got a new data data with the real data for the period from 1^{st} January 2008 to 31^{st} December 2010, further named real data.

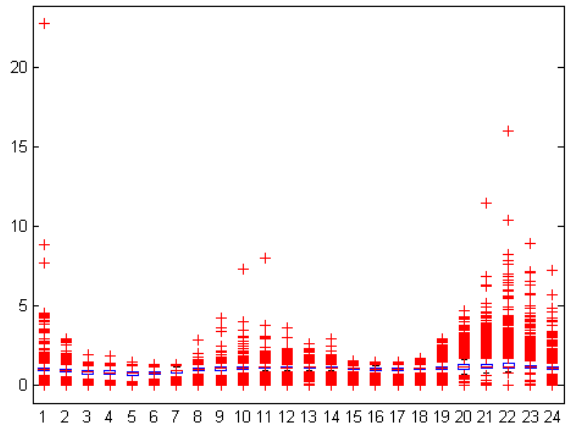


Figure 5: Boxplot diagrams (rescaled data 2008-2016).

Next, the initial exploratory data analysis more significant to each data set previously described (rescaled data and real data), is presented. The boxplot diagrams (Figure 5) concerning rescaled data (2008-2016) reveal different distributions and a great number of anomalies per hour.

Since we have a huge dimensional dataset, we restrict the presented graphics in Table 1 to the year 2010 to compare the rescaled dataset and the real data set. From Table 1 we can conclude that rescaled data present a huge quantity of "uncommon" observations each hour of the day with exception of hours 4, 5 and 6. The rescaled data also have different patterns off dispersion. By other hand, the real data display unusual observations but in a fewer quantity than rescaled data. The dispersion of real data presents more homogeneous patterns each hour. These details can be confirmed in next table, where are summarized some descriptive statistics and tests.

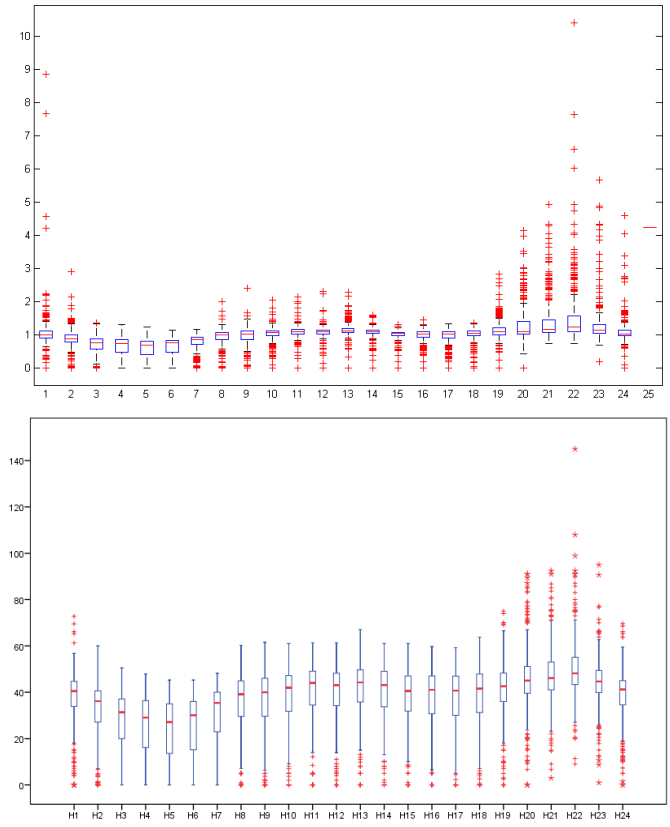


Table 1: Boxplot diagrams of rescaled (*top*) and 2010 real data (*bottom*).

From Table 2, we can see the different patterns of dispersion (observe the standard deviation and inter-quartile range columns respectively). Also we confirm that the data does not have normal distribution when we check the

Hora	mean	trimmean	median	std	iqr
1.0000	1.0212	0.9908	0.9900	0.5327	0.1800
2.0000	0.8695	0.8872	0.8900	0.2487	0.2000
3.0000	0.7531	0.7943	0.8100	0.2483	0.2100
4.0000	0.7114	0.7541	0.7800	0.2523	0.2300
5.0000	0.6802	0.7230	0.7500	0.2504	0.2400
6.0000	0.7107	0.7573	0.7800	0.2369	0.1900
7.0000	0.8111	0.8594	0.8700	0.2211	0.1600
8.0000	0.9488	0.9773	0.9900	0.2067	0.1600
9.0000	0.9911	1.0163	1.0300	0.2457	0.1900
10.0000	1.0582	1.0666	1.0700	0.2596	0.1500
11.0000	1.0975	1.0988	1.1000	0.2322	0.1200
12.0000	1.0823	1.0896	1.0900	0.1724	0.1100
13.0000	1.0955	1.0998	1.1000	0.1633	0.1200
14.0000	1.0709	1.0807	1.0800	0.1597	0.1100
15.0000	1.0096	1.0282	1.0300	0.1575	0.1000
16.0000	0.9690	0.9973	1.0000	0.1774	0.1300
17.0000	0.9547	0.9872	1.0000	0.1913	0.1500
18.0000	0.9987	1.0209	1.0300	0.1843	0.1400
19.0000	1.0861	1.0715	1.0600	0.2388	0.1700
20.0000	1.1944	1.1275	1.1000	0.3818	0.2500
21.0000	1.2651	1.1717	1.1500	0.4885	0.2575
22.0000	1.3302	1.2027	1.1700	0.6446	0.2400
23.0000	1.2139	1.1298	1.1100	0.4657	0.2000
24.0000	1.0760	1.0336	1.0200	0.3265	0.1700

Hora	skewness	kurtosis	P-value (KS)	P-value (JB)
1.0000	24.5154	927.2962	0	0.0010
2.0000	-0.0747	12.7788	0	0.0010
3.0000	-1.4003	5.3185	0	0.0010
4.0000	-1.2967	4.6556	0	0.0010
5.0000	-1.2724	4.2777	0	0.0010
6.0000	-1.6251	5.3662	0	0.0010
7.0000	-2.2851	8.3236	0	0.0010
8.0000	-1.8786	12.7847	0	0.0010
9.0000	0.0254	24.4068	0	0.0010
10.0000	4.8500	122.5760	0	0.0010
11.0000	8.7122	269.0530	0	0.0010
12.0000	-0.2557	39.8107	0	0.0010
13.0000	-1.5259	24.3516	0	0.0010
14.0000	-1.8895	28.5563	0	0.0010
15.0000	-3.4210	21.3683	0	0.0010
16.0000	-2.9380	15.0985	0	0.0010
17.0000	-2.6087	12.1901	0	0.0010
18.0000	-2.4575	13.7446	0	0.0010
19.0000	0.8480	13.3137	0	0.0010
20.0000	3.1875	20.0743	0	0.0010
21.0000	6.6494	88.1468	0	0.0010
22.0000	8.4558	126.6330	0	0.0010
23.0000	7.0455	75.4992	0	0.0010
24.0000	6.5014	82.1613	0	0.0010

Table 2: Descriptive summary (rescaled data 2008-2016). Left: Mean, trimmean, media, standard deviation, inter-quartile range. Right: Skewness, kurtosis, Kolmogorov-Smirnov, and Jarcke and Bera normality tests.

Kolmogorov-Smirnov and Jarcke and Bera normality tests.

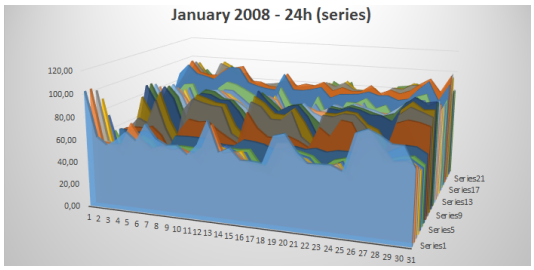


Figure 6: Real data (January 2008).

Considering the real data, for example from January 2008, we found different patterns per day and per hour (see Figure 6). The same behaviour was found in Figure 7, where, for example, we can see that 22 groups (hours) have mean ranks significantly different from group 1 (hour 1).

3 GLM Approach

In 1972, had borned the idea of GLM as a powerful method in Statistics, standardizing the theoretical and applied points of view about all the structure of linear regression developed until that time. Due to the large number

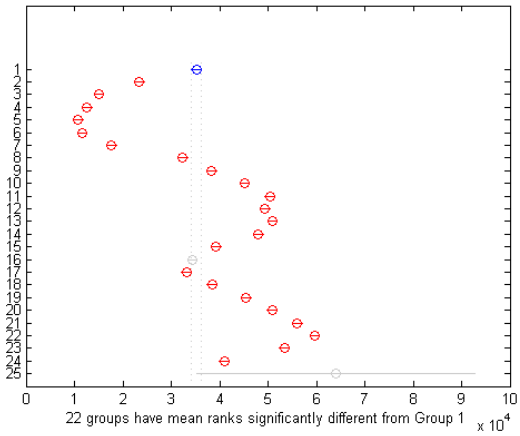


Figure 7: Real data (January 2008). Mean price per hour.

of models, and simplicity of development associated with rapid computational analysis, the GLM have been playing an important role in statistical analysis. The idea of GLM is the establishment of a functional relation between the variable to predict (dependent variable) and a set of other exogenous variables (explanatory variables or covariates). This relation allows to predict the dependent variable. The dependent variables and the explanatory variables can be of any type: continuous, discrete, dichotomous, quantitative, qualitative, stochastic, non-stochastic. The response variable can also be a proportion, be positive, have a non-normal random component. At 1935, Bliss proposed the probit model to proportions; in 1944 Berkson developed the logistic regression, log-linear models for contingency tables were introduced by Birch at 1963. In 1972, Nelder and Wedderbrun proved that all these models are particular cases of a general family: the generalized linear models. In GLM, the random component belongs to exponential family[1] and a transformation of expected value of response variable is related with explanatory variables. The simplest models, where the explanatory variables are nonrandom and the disturbances are gaussian white noise, which are estimated by ordinary least squares, can be extended for more general models in which the disturbances are autocorrelated, heteroscedastic, not gaussian, etc, or when some of the explanatory variables are stochastic. So,

[1]One random variable Y belongs to exponentially distributed family if it probability density function (or probability mass function) can be represented as

$$f(y|\theta,\phi) = e^{\frac{y\theta - b(\theta)}{a(\theta)}} + c(y,\phi), \tag{1}$$

where θ and ϕ are parameters, $a(.)$, $b(.)$ and $c(.)$ are known functions that verify some regularity conditions. The parameters and θ and ϕ are the canonical form of localization and dispersion parameters respectively. If ϕ is known, f belongs to uni-parametric exponential family, if ϕ is unknown, f belongs to bi-parametric exponential family.

linear regression models can be estimated by generalized least squares.

In the classical linear model, a vector X with p covariates $X = (X_1, X_2, \ldots, X_p)$ can explain the variability of the variable of interest Y (response variable), where $Y = Z\beta + \epsilon$. Z is a specification matrix with size $n \times p$ (usually $Z = X$, considering an unitary vector in first column), β a parameter vector and ϵ a vector of random errors ϵ_i, independent and identical distributed to a reduced gaussian.

The data are in the form (y_i, x_i), $i = 1, \ldots, n$, as result of observation of (Y, X) n times. The response variable Y has expected value $E[Y|Z] = \mu$.

In GLM, the model is an extension of classical model, the response variable, following an exponential family distribution [7], do not need to be gaussian.

Another extension from the classical model is that the function which relates the expected value and the covariates can be any differentiable function. Y_i has expected value $E[Y_i|x_i] = \mu_i = b'(\theta_i)$, $i = 1, \ldots, n$.

It is also defined a differentiable and monotone link function g which relates the random component with the systematic component of response variable. The expected value μ_i is related with the linear predictor $\eta_i = z_i^T \beta_i$ using the relation

$$\mu_i = h(\eta_i) = h(z_i^T \beta_i), \qquad \eta_i = g(\mu_i) \tag{2}$$

where

- h is a differentiable function;
- $g = h^{-1}$ is the function link ;
- β is a vector of parameter with size p (the same size of the number of explanatory variables);
- Z is a specification vector with size p.

There are different link functions in GLM. When the random component of response variable has a Poisson distribution, the link function is logarithmic and the model is log-linear. In particular, when the linear predictor $\eta_i = z_i^T \beta_i$ coincides withe the canonical parameter θ_i, $\theta_i = \eta_i$, which implies $\theta_i = z_i^T \beta_i$, the link function is denominated as canonical link function.

The GLM methodology can be summarized in three steps:

1. Models formulation: identify response variable distribution, select the preliminaries covariates and specification matrix, select the link function g;
2. Models adjustment: estimation of model parameters, application of suitability measures of estimates;

3. Selection and validation of models: selection of variables, diagnostics, residual analysis and interpretation.

Initially, the GLM approach using IBM SPSS Statistics (version 20) could not be performed due the high dimensionality of data. To solve partially such issue, we try to reduce the 24 hours of a day to fewer reference time intervals. By first, we analyse the data plot per hour. The graphical representation of data, see Figure 3, shows similar behavior in some distinct. Identified such similar hours we merge them in an unique interval of similarity. In this way the dimension of data can be reduced, by taking the mean or median or other measure of response variable.

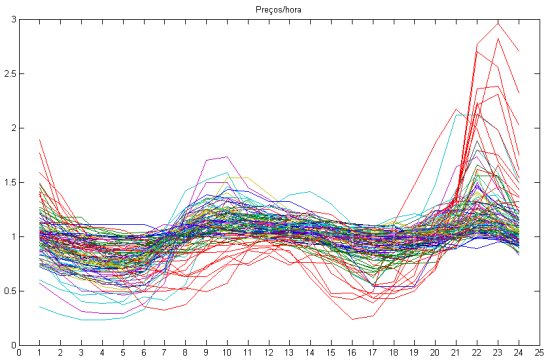


Figure 8: Data representation

Following the described methodology, we have selected and defined some time intervals which conduced to the best model performance. In this way, we have reduced the dimension defining the following time intervals: aurora, lunch time and dinner time. Aurora corresponds to the hours 3, 4 and 5 respectively. Lunch time merges the hours 11, 12, 13 and 14. Dinner time takes into account hours 17, 18 and 19. We have overlaped the data graphically, for each hour in the defined time intervals (see Figures 9, 10, 11). We do not notice significant differences.

We studied all possible covariates which can contribute to the explication of price per hour. Using the initial covariates proposed in Section 2 and defining the codification configuration $X_1 = C_1$, $X_2 = C_4$, $X_3 = C_5$, $X_4 = C_6$, $X_5 = C_7$, it was performed one analysis of variance (ANOVA) with second order interaction in a preliminaire stage of the study, being chosen the best candidates to covariables of a GLM model. The results of such a analysis is presented in Table 12.

It was also considered the fare defined by EDP (see Figure 13) as possible covariate but it was not an important explanatory variable.

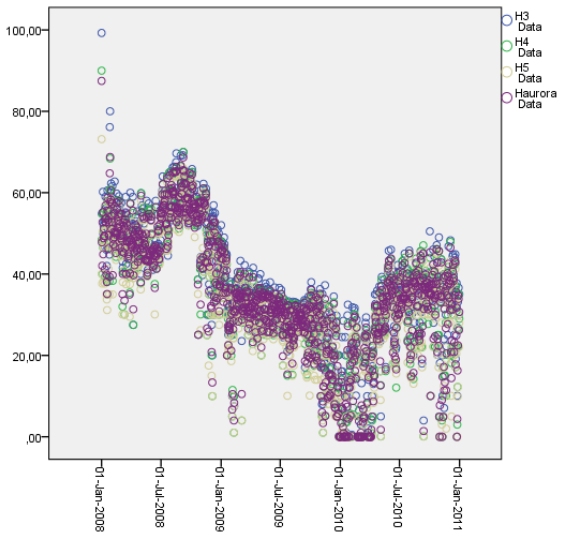


Figure 9: Olerlaped data: Aurora.

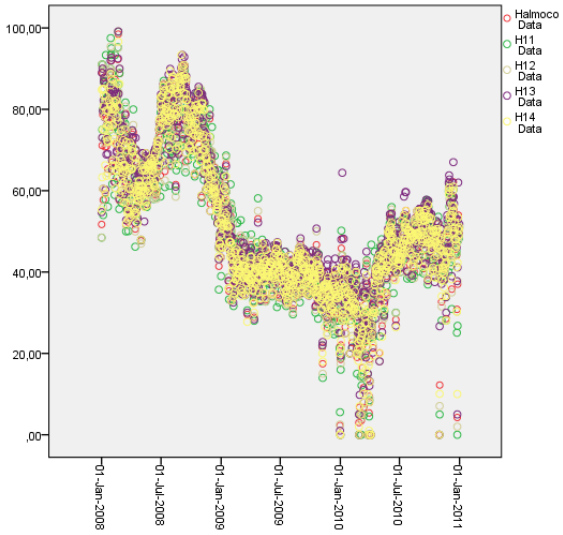


Figure 10: Olerlaped data: Lunch time.

When we estimate the best models, after the selection stage, the dependent variable was transformed by log or square root. The significant covariates were C4, C6, C7, H2, H7, H8, H16, H20, H22, H23, H24 and lunch time (square root transformation) and C4, C6, C7, H2, H7, H8, H16, H20, H22, H23, H24 (log transformation). Notice that other transformations should be considered taking into account the time series nature of the

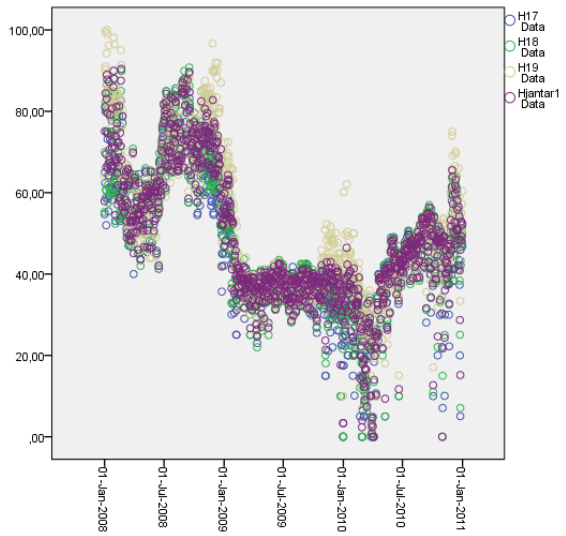


Figure 11: Olerlaped data: Dinner time.

Analysis of Variance

	Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
	X1	89.38	23	3.88601	175.44	0
	X2	0.74	6	0.12359	5.58	0
#	X3	0	0	0	0	NaN
#	X4	0	0	0	0	NaN
#	X5	0	0	0	0	NaN
#	X6	0	0	0	0	NaN
	X1*X2	132.75	138	0.96197	43.43	0
	X1*X3	13.94	23	0.60618	27.37	0
	X1*X4	8.41	69	0.12194	5.51	0
	X1*X5	35.13	253	0.13887	6.27	0
	X1*X6	1.08	23	0.04699	2.12	0.0013
	X2*X3	1.1	6	0.18269	8.25	0
	X2*X4	0.83	18	0.04617	2.08	0.0045
	X2*X5	2.74	66	0.04147	1.87	0
	X2*X6	0.52	6	0.08713	3.93	0.0006
#	X3*X4	0.35	2	0.17549	7.92	0.0004
#	X3*X5	5.92	9	0.65833	29.72	0
	X3*X6	2.86	1	2.86129	129.17	0
#	X4*X5	0.07	2	0.03306	1.49	0.2248
#	X4*X6	0	0	0	0	NaN
#	X5*X6	0.01	1	0.00888	0.4	0.5267
	Error	1632.52	73701	0.02215		
	Total	2664.46	74366			

Constrained (Type III) sums of squares. Terms marked with # are not full rank.

Figure 12: ANOVA with secondary effects.

data. This way we could probably get better models.

Considering the obtained results as indicators, we can conclude that some of the covariates proposed initially were not relevant for dependent variable, such as, EDP fares, Portuguese holidays (maybe the Iberian holidays can have some relevance, and not just the Portuguese ones). Also, some periods of time can be drop off as relevant covariates, such as dinner time or some others. The season, month or winter/summer time period revealed

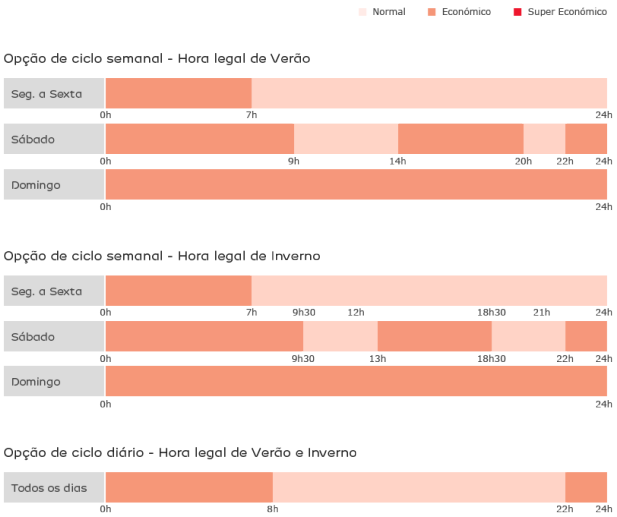


Figure 13: EDP fares in https://www.edp.pt.

significant explanatory variables in the different estimated models.

4 Multivariate Models

In this section a multivariate time series approach is applied to the data correspondent to the hourly prices from 01/01/2014 to 28/06/2016[2]. The data analysis presented in this Section was performed using RStudio (version 0.99.9902) and R Statistical Software (version 3.3.0)[10].

Figure 14 shows the MIBEL daily prices of the 24 hours from 01/01/2014 to 28/06/2016, in a total of 910 observations. The data from 30/03/2014, 29/03/2015 and 27/03/2016 only concern 23 legal hours, therefore, these missing value was filled with the previous value with the assumption that the current data will be similar to the previous ones (see also [1]). Furthermore, there were several zero hourly prices in the first observations of this time series, more precisely on the first 68th days in 2014. For this reason the first 68 observations were removed and therefore a total 842 observation were considered. Table 3 shows the mean, standard deviation, median and range for each the 24 hourly prices, Y_{kt}, $k = 1, ..., 24$, after this first treatment to the data.

For stabilizing the variance the log transformation was applied (see Figure 15). The existence of linear dynamic dependence in the data is supported by the multivariate LjungBox test (not presented here for simplicity). For

[2]For a discussion on fitting models to short time series see e.g. http://robjhyndman.com/hyndsight/short-time-series/

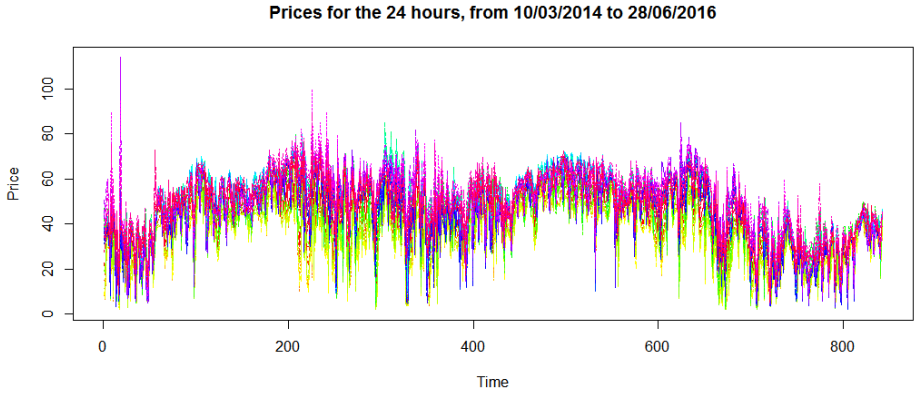


Figure 14: MIBEL hourly prices for each of the 24 hours from 01/01/2014 to 28/06/2016 (data provided by EDP during the 119th ESGI).

Variable	mean	sd	median	range	Variable	mean	sd	median	range
H1	42.63	13.62	44.37	64.91	H13	48.76	15.21	51.27	76.06
H2	38.15	13.32	40.48	66.20	H14	48.29	15.14	50.50	74.55
H3	34.94	13.27	38.14	60.38	H15	45.97	15.06	48.17	70.00
H4	34.02	13.44	36.86	59.80	H16	43.87	15.34	45.40	68.71
H5	33.22	13.47	35.44	59.90	H17	43.52	15.73	45.11	70.53
H6	34.48	13.20	37.00	59.88	H18	45.02	15.67	46.26	70.54
H7	38.81	13.65	40.97	64.70	H19	47.29	15.40	48.82	73.69
H8	43.92	14.69	45.37	70.91	H20	49.78	14.79	51.01	107.92
H9	46.76	14.89	48.45	73.60	H21	51.89	14.08	54.03	92.58
H10	49.01	14.98	51.32	75.49	H22	53.06	13.06	55.98	79.50
H11	49.71	14.92	52.00	81.05	H23	50.35	12.92	53.20	68.53
H12	49.20	15.00	51.98	75.50	H24	45.62	12.85	47.00	66.30

Table 3: Mean, standard deviation, median and range for the 839 observations of each the 24 hourly prices. Here Hk stands for Y_{kt}, $k = 1, \ldots, 24$.

the Augmented Dickey-Fuller Test one may assume that the time series are stationary (see Table 4).

In order to explore the intra-day correlation structure, the hourly spot prices, $\boldsymbol{Y}_t = (Y_{1,t}, \ldots, Y_{24,t})^\top$, can be represented as a Vector Autoregressive model of order p, VAR(p):

$$\boldsymbol{Y}_t = \phi_0 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{Y}_{t-i} + \boldsymbol{a}_t, \tag{3}$$

where ϕ_0 denotes the constant vector, $\boldsymbol{\Phi}_i$ are 24×24 matrices of autoregressive parameters and $\boldsymbol{a}_t$ the residuals. This approach allows to model the hourly spot prices jointly and therefore capture the correlation between the hourly price, as has been reported in previous studies (see e.g. [13]). The drawback is that this approach requires the estimation of $24 \times (24 \times p + 1)$.

	ADF	p		ADF	p
1	-4.65	<0.01	13	-3.89	0.01
2	-5.00	<0.01	14	-3.97	0.01
3	-5.10	<0.01	15	-4.13	< 0.01
4	-4.94	<0.01	16	-4.37	< 0.01
5	-4.86	<0.01	17	-4.25	< 0.01
6	-4.91	<0.01	18	-3.92	0.01
7	-4.98	<0.01	19	-3.49	0.04
8	-4.70	<0.01	20	-3.31	0.07
9	-4.47	<0.01	21	-3.27	0.08
10	-4.21	<0.01	22	-3.44	0.05
11	-3.91	<0.01	23	-3.54	0.04
12	-3.84	0.02	24	-3.89	0.01

	ADF	p		ADF	p
1	-4.96	< 0.01	13	-4.13	< 0.01
2	-5.49	< 0.01	14	-4.21	< 0.01
3	-5.56	< 0.01	15	-4.39	< 0.01
4	-5.39	< 0.01	16	-4.68	< 0.01
5	-5.27	< 0.01	17	-4.53	< 0.01
6	-5.41	< 0.01	18	-4.19	< 0.01
7	-5.38	< 0.01	19	-3.74	0.02
8	-5.00	< 0.01	20	-3.50	0.04
9	-4.71	< 0.01	21	-3.40	0.05
10	-4.50	< 0.01	22	-3.57	0.04
11	-4.07	< 0.01	23	-3.64	0.03
12	-4.03	< 0.01	24	-4.02	< 0.01

Table 4: Results for the stationarity tests without the logarithm transformation (*on the left*) and after the transformation (*on the right*). Recall that for the Augmented Dickey-Fuller the alternative hypothesis is stationarity.

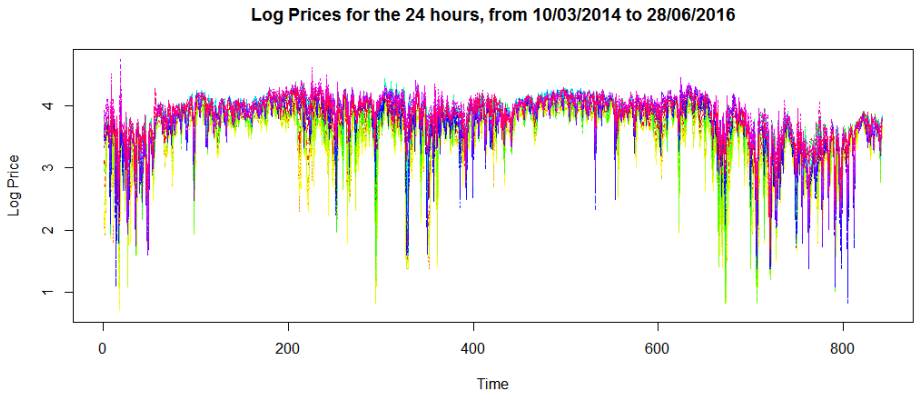


Figure 15: MIBEL Log hourly prices for each of the 24 hours from 10/03/2014 to 28/06/2016 (data provided by EDP during ESGI119).

To capture the seasonal pattern of the process mean we add to it exogenous variables, $\boldsymbol{X}$, yielding a VARX model (see [12]). The general form of a VARX of orders p and s is:

$$\boldsymbol{Y}_t = \phi_0 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{Y}_{t-i} + \sum_{j=0}^{s} \boldsymbol{\beta}_j \boldsymbol{X}_{t-j} + \boldsymbol{a}_t. \tag{4}$$

A similar model, with $p = 7$, was presented by [8], although they consider $\boldsymbol{\Phi}_i$ as diagonal matrices. In that work the authors use a dummy representing the day type (working day vs. weekend) and the number of daylight hours (which mimics the annual seasonality).

We use as exogenous two variables: day type (working day and weekend)

and the annual seasons defined in terms of meteorological conditions[3].

A VARX$(7,0)$ was estimated using hourly prices Y_{kt}, $k = 1, ..., 24$, from 10/03/2014 to 29/05/2016, in a total of 811 observations. Note that $p = 7$ is chosen in several of the EPF studies presented in literature. The prices from 30/05 to 28/06/2016 were used for comparing the real hourly prices with the forecast of this model. A total of 4104 $(= 24+24\times24\times7+24\times2)$ parameters were estimated using functions provided by the `MTS` package from R. The full model was reduced by removing simultaneously all estimates with t-ratio less than 0.5. From Figure 16 we can see that the model is adequate. The p-values of the Ljung-Box statistics are all > 0.05, and therefore the null hypothesis of zero cross-correlations is not rejected.

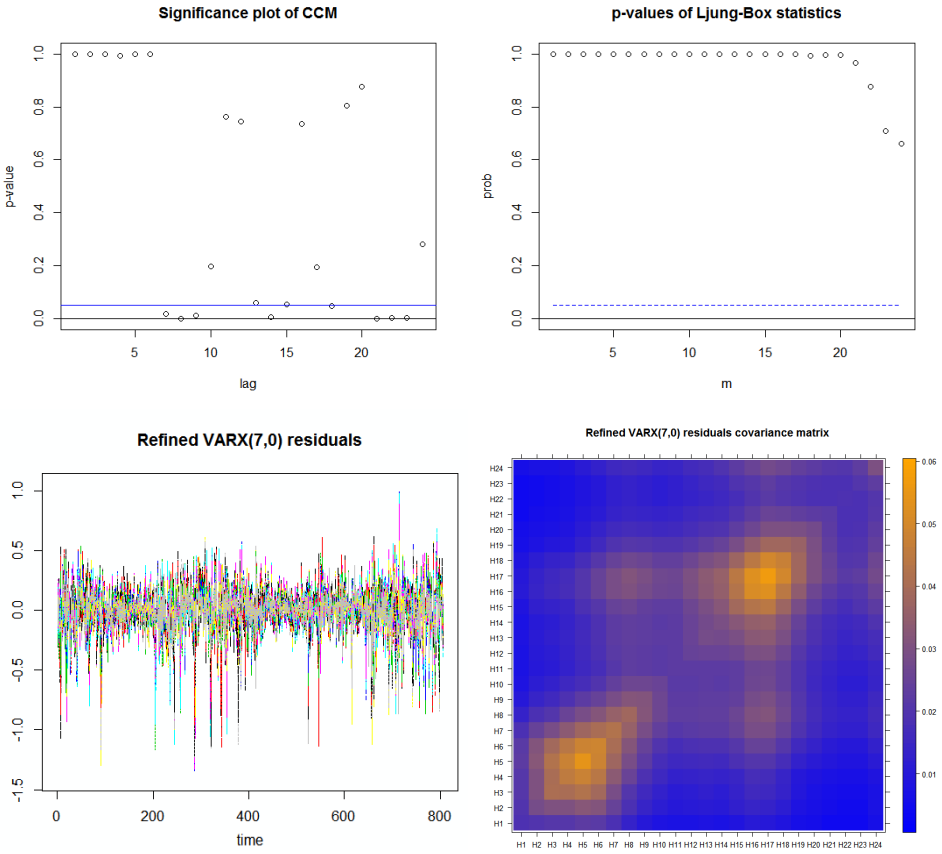


Figure 16: VARX(7,0) residual analysis.

The autoregressive coefficients of the VARX(7,0), explain the existence (or not) of dependence within the hourly prices. The element (k, j) with $k \neq j$ of the $\mathbf{\Phi}_i$ matrices shows the linear dependence of Y_{kt} on $Y_{j,t-1}$ in

[3]See http://www.calendario-365.pt/epocas-estacoes-do-ano.html

the presence of $Y_{k,t-1}$. Figure 17 shows that there are several non-diagonal elements of Φ_i that are nonzero, therefore there are several hourly prices that are dynamically correlated to prices of other hours in previous days.

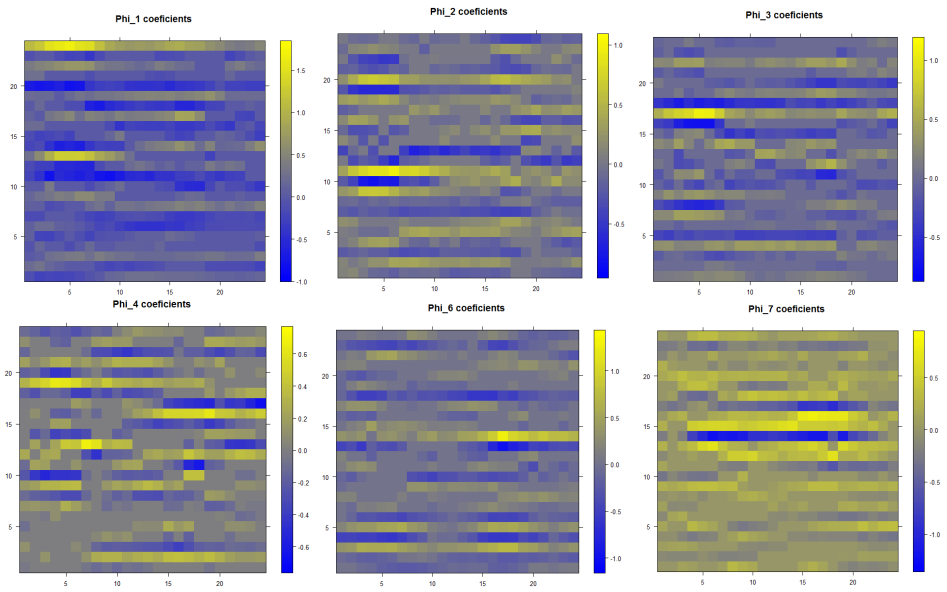


Figure 17: Autoregressive coefficients of the VARX(7,0). For simplicity the coefficients of $\mathbf{\Phi}_5$ were omitted.

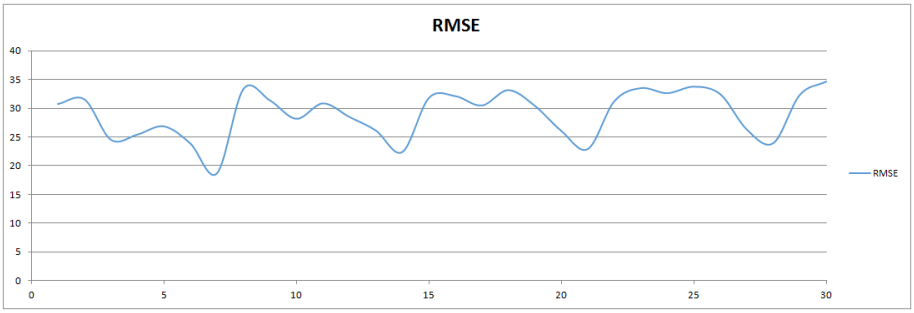


Figure 18: RMSE for the forecasts from 30/05 to 28/06 using the reduced VARX(7,0).

For evaluating the point forecasts the root mean square errors (RMSE) was used. According to [13] this is perhaps the most popular measure for this purpose. Figure 18 presents the RMSE for the 30 days forecast. The forecasts were obtained from a VARX(7,0) estimated using log of the hourly prices. These forecasts were then transform back by applying exp. As an

example, in Figure 19 the real hourly prices and forecasts are depicted for the first four days. Although the forecast do not exactly replicate the real price they are quite similar. For example the introduction into the model of exogenous variables that better explain the meteorological condition would certainly improve the model. On the other and further analysis of the orders of the VARX should be performed.

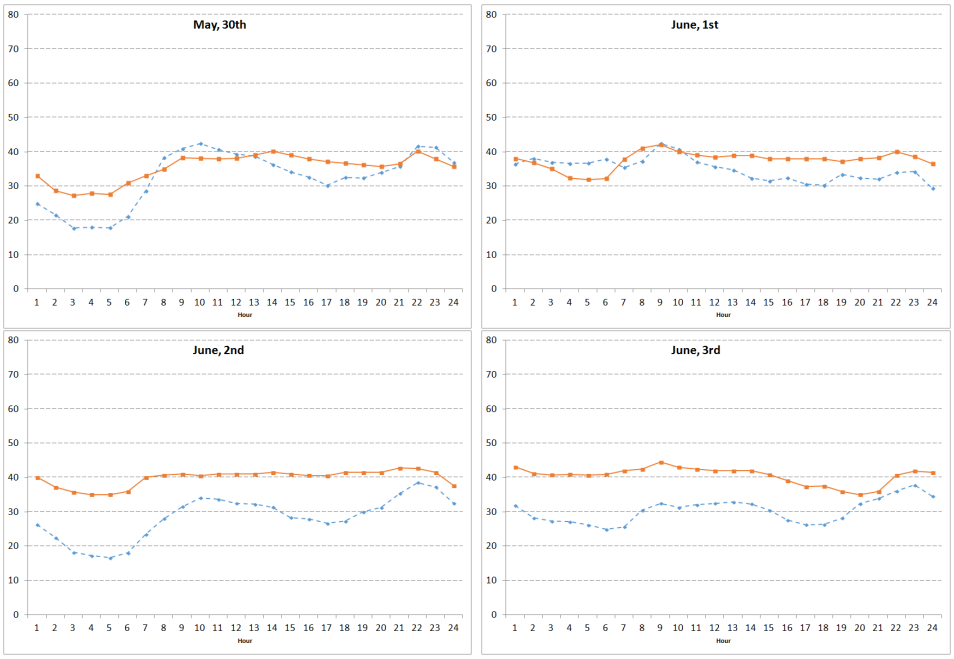


Figure 19: Real (*orange full line*) and forecast (*blue dot line*) hourly prices using VARX(7,0)

5 Conclusions and recommendations

The challenge proposed by EDP consisted in simulating electricity prices not only for risk measures purposes but also for scenario analysis in terms of pricing and strategy. Data concerning hourly electricity prices from 2008 to 2016 were provided by EDP.

One week is certainly not enough to work on this challenge. During that week this study group pointed out different promising statistical techniques and possible approaches, namely: ARIMA, sARIMA, Longitudinal Models, Generalized Linear Models and Vector Autoregressive Models.

The data was explored using different statistical software, namely IBM SPSS Statistics, Matlab and R Statistical Software. In this report a GLM and a vector autoregressive modelwere considered. In the GLM framework

two different transformations were consider and for both the season of the year, month or winter/summer period revealed significant explanatory variables in the different estimated models.

On the other hand the multivariate approach using VAR considering as exogenous variables the meteorologic season and the type of day yield a multivariate model that explains the intra-day and intra-hour dynamics of the hourly prices. Although the forecast do not exactly replicate the real price they are quite similar. We believe that introducing exogenous variables that better explain the meteorological conditions would certainly improve the model and the forecast. On the other, further analysis of the orders of the VARX should be performed.

In both of these approaches a more extensive work would certainly improve the proposed models.

However, others approaches should be explored. Univariate time series are one of these approaches. Although several authors have presented models using different univariate approaches, topics such as diagnostic analysis and selection of the order of the models seems to be forgotten, or at list put aside. On the other hand, to our knowledge, longitudinal modeling have not yet been addressed in Electricity Price Forecasting (EPF), and is an approach that we consider deserves further attention.

In conclusions, EPF is a growing area that groups multiple different approaches that can be applied. In fact, other approaches from multi-agent models, fundamental models, reduced-form models and computational intelligence models, also present a great space for EPF.

References

[1] P. Chujai, N. Kerdprasop, and K. Kerdprasop. Time series analysis of household electric consumption with arima and arma models. In *Proc. IMECS Conf., Hong Kong*, 2013.

[2] P. Diggle, P. Heagerty, K.Y. Liang, and S. Zeger. *Analysis of Longitudinal Data*. Wiley, second edition, 2002.

[3] S. S. Joens et al. A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of Biomedical Informatics*, 42:123–139, 2009.

[4] Y. Fu et al. Arfnns with svr for prediction of chaotic time series with outliers. *Expert Systems with Applications*, 37:4441–4451, 2014.

[5] B. Poczos et. all. Nonparametric kernel estimators for image classification. Technical report, 2012.

[6] J. W. Taylorł J. M. Jeon. Using conditional kernel density estimation for wind power density forecasting. *J. American Statistical Association*, 107, 2012.

[7] G. Silva M. A. Turkman. *Modelos Lineares Generalizados da teoria a prática*. Sociedade Portuguesa de Estastística, Lisboa, 2000.

[8] K. Maciejowska and R. Weron. Forecasting of daily electricity prices with factor models: utilizing intra-day and inter-zone relationships. *Computational Statistics*, 30(3):805–819, 2015.

[9] G.G.P. Murthy, V. Sedidi, A.K. Panda, and B.N. Rath. Forecasting electricity prices in deregulated wholesale spot electricity market-a review. *International Journal of Energy Economics and Policy*, 4(1):32, 2014.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[11] L. Su. Prediction of multivariate chaotic time series with local polynomial fitting. *Computers & Mathematics with Applications*, 59(2):737–744, 2010.

[12] R. S. Tsay. *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons, 2014.

[13] R. Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4):1030–1081, 2014.