Open Journal of Statistics, 2019, 9, 230-244
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Constructing Prediction Intervals: A Re-Edition of the Williams-Goodman Method

Ilaria L. Amerise*, **Agostino Tarsitano**

Dipartimento di Economia, Statistica e Finanza, Università della Calabria via Pietro Bucci, Rende, Italy
Email: *ilaria.amerise@unical.it, agostino.tarsitano@unical.it

How to cite this paper: Amerise, I.L. and Tarsitano, A. (2019) Constructing Prediction Intervals: A Re-Edition of the Williams-Goodman Method. *Open Journal of Statistics*, **9**, 230-244.
https://doi.org/10.4236/ojs.2019.92017

Received: February 22, 2019
Accepted: April 16, 2019
Published: April 19, 2019

Abstract

The aim of this paper is to develop and validate a procedure for constructing prediction intervals. These forecasts are produced by Box-Jenkins processes with external deterministic regressors and prediction intervals are based on the procedure proposed by Williams-Goodman in 1971. Specifically, the distributions of forecast error at various lead-times are determined using post-sample forecast errors. Fitting a density function to each distribution provides a good alternative to simply observing the errors directly because, if the fitting is satisfactory, the quantiles of the distribution can be estimated and then the interval bounds computed for different time origins. We examine a wide variety of probability densities to search the one that best fit the empirical distributions of forecast errors. The most suitable mathematical form results to be Johnson's system of density functions. The results obtained with several time series suggest that a Box-Jenkins process combined with the Williams-Goodman procedure based on Johnson's distributions, provide accurate prediction intervals.

Keywords

Electricity Market, Post-Sample Errors, Probabilistic Forecast, Zonal Prices

1. Introduction

Energy companies are strongly affected by uncertain price conditions, as they are exposed to the different risks from liberalized energy markets in combination with important and, to a large extent, irreversible investments. Price predictions, however, are usually expressed as point forecasts that give little guidance as to their accuracy, whereas, the planning process needs to take into account the entire probability distribution of future prices or at least intervals that have a pre-specified nominal coverage rate *i.e.* a given probability of containing the future prices. It is the aim of this paper to resume the prediction intervals (PIs)

proposed by Williams & Goodman [1], which anticipated bootstrap techniques, but was introduced at a time when its heavy computational demands slowed down practical applications.

A literature review in the field of short-term forecasting of electricity prices, reveals that limited research has addressed the issue of PIs. Misiorek *et al.* and Weron [2] [3] were the first to consider interval forecasts. Other few exceptions are Wu *et al.*, Nowotarski & Weron, and Bunn *et al.* [4] [5] [6]. However, interval forecasts remain an underdeveloped topic. See [7]. Under the very restrictive assumptions of independent and identically distributed Gaussian errors, PIs can be obtained by means of standard formulae. Since the distribution of forecast errors is unknown, they must be estimated to calculate interval bounds.

The main problem with assessing the reliability of price forecasts is that the magnitude of post-sample errors cannot be exactly evaluated until the prices are observed. In order to simulate such a situation, the time series under study can be split into two parts: the "training" period, which ignores a number of the most recent time points and the "validation" period, which comprises only the ignored time points. For the purpose of this study, we have not used the entire time series, but kept the very last time points untouched because they serve as a benchmark (target period) against which the quality of the PIs is to be judged. The training period is used to identify and estimate one of the large variety of electricity price models described in literature. Borovkova & Schmeck [8] observe that, despite the voluminous literature on modeling electricity prices, a clear "winner" model has not emerged. Here, we use a Box-Jenkins process.

Williams & Goodman [1] had the simple and ingenious idea of rolling ahead the training period and repeating the forecasting procedure until it is not possible to make any more multi-step predictions. The collection of multi-step-ahead errors forms the empirical distribution of the forecast errors for each lead-time. This allows us to estimate the quantiles necessary to compute the interval bounds. To implement the Williams & Goodman (WG) procedure, it is necessary to find a family of probability distributions having a member which gives a good fit to the empirical distribution of forecast errors. We will consider various density functions to be combined with SARIMAX processes for time series data concerning Italian electricity hourly zonal prices. Our aim is to find the most effective way of mixing WG procedure and SARMAX processes.

The organization of the paper is as follows. In the next section we address important aspects of data-preparation. Section 3 provides a brief review of the Box Jenkins approach to compute point forecasts. In Section 4, the Williams-Goodman procedure is discussed and in Section 5 it is combined with various density functions purportedly useful in describing the empirical forecast error distribution. In this same section is presented an application to Italian hourly zonal prices. Conclusions are drawn in Section 6.

2. Data Preparation

In this article we analyze data on hourly zonal prices traded at the day ahead

Italian energy market. Because of transmission capacity constraints, Italy is partitioned into six zones: North, Centre-North, Centre-South, South, Sardinia and Sicily with a separate price for each zone. When there are no transmission congestions, arbitrage opportunities restrict the prices in each zone to be equal. See [9]. The hourly marginal prices can differ across zones because of transmission limits or because of dissimilar behavior of the consumers, but it is the same within a zone. The sizes of the zones are not equal. In fact, North (which includes the Italian regions of Val D'Aosta, Piemonte, Liguria, Lombardia, Trentino, Veneto, Friuli Venezia Giulia, Emilia Romagna) constitutes a large fraction of Italian national surface (40%) and its population (46%). The large islands of Sicily and Sardinia suffer from poor interconnections and frequent congestions. In the present article, we prefer to work with zonal data because they show far wilder randomness than the national price and are complex enough to challenge the statistical methods used for making predictions.

Data sets are freely accessible by the Italian independent system operator on http://www.mercatoelettrico.org/En/Tools/Accessodati.aspx?ReturnUrl=%2fEn%2fDownload%2fDatiStorici.aspx.

According to principles of decentralization and subsidiarity, creatively extended to long time series, we will treat each hour as a separate time series, such that 24 different models are estimated. All the time series go from 1 am on Monday, 7/1/2013 to 24 pm on Sunday, 26/2/2017 and hence cover a total of 24 hourly prices in 1148 days for six zones. We reconstructed the values corresponding to the changes of the daylight-saving time by the arithmetic average of the two neighboring hours, while the "doubled" values corresponding to the switch from the daylight-saving time, were replaced by the arithmetic mean of the two neighboring prices.

Time series of electricity prices display characteristics not frequently observed in other commodity markets. Pronounced daily, weekly, monthly and multi-monthly seasonal cycles; heteroskedasticity, mean reversion, and a high number of spikes (very sharp peaks or extremely deep valleys) in a short period of time. Knittel & Roberts [10] note that electricity prices also contain what they refer to as, an "inverse leverage effect". Electricity price volatility tends to rise more so with positive shocks than negative shocks. Most of these characteristics stem from the fact that power cannot be economically stored and, consequently, accumulation and selling of stocks/inventories have reduced intervention potential to smooth supply or demand shocks across over time. Additionally, electricity markets face stringent distribution and transmission constraints.

To attenuate these effects, prices are log-transformed so that all upward or downward spikes are closer to the mean of the time series. The attenuation does not absolve us from trying to use more effective albeit more invasive treatment of aberrant prices. On the one hand, even if the removal of legitimate data points could be accepted as a permissible practice, the number of values suspectable of being anomalous is too large to justify their exclusion. Extreme price swings, in fact, need not be treated as enemies, because they are very significant for energy

market participants. On the other hand, as noted by Fildes [11], the choice of a forecasting method should not be dependent on such extremes, unless they contain information we cannot afford not to consider.

To deal with spike prices, we construct an artificial time series by decomposing the original time series into trend-cycle (expressed through orthogonal polynomials) and periodicities (expressed as a sum of harmonics with random phases). Deviations between observed and artificial prices outside the range: median plus/minus a factor of the median absolute deviation from the median, are considered anomalous residuals which may indicate abnormal price. These prices, are considered missing and replaced by a weighted average of the observed prices and the corresponding artificial prices. Although infrequent, negative or virtually zero prices do occur. These unusual prices can create problems with log transformation. So, prices less than one e/MWh are treated as missing values and imputed using the artificial time series.

3. Point Forecast

The generic time series is represented by $P_1, P_2, \cdots, P_n$ where n is the number of observations, which, in this section, comprises the training and the validation periods (which, taken together, form the fit period). The index of the hour is suppressed, but it is understood that P_t refers to a daily time series of one of the hours.

There is no general consensus at present on the best method to be used for electricity prices modeling. In this context, we apply Box-Jenkins forecasting method which has proven to be flexible enough to accommodate the electricity price behavior satisfactory. See [12] [13]. We do not investigate other time series approaches which are potentially useful, but lie beyond the scope of the present study. For a recent comprehensive survey see [14].

The general form of a sarimax model is

$$P_t = \beta_0 + \sum_{j=1}^{m} \beta_j X_{t,j} + \left[\phi^*(B)\right]^{-1} \theta^*(B) a_t, \tag{1}$$

where P_t is the price at the day t and a_t is a white noise process with zero mean and finite variance σ_a^2. The symbol B represents the usual backward shift operator and $\phi^*(B)$ and $\theta^*(B)$ are polynomials in B

$$\begin{cases} \phi^*(B) = 1 - \phi_1^* B - \phi_2^* B^2 - \cdots - \phi_{p^*}^* B^{p^*} \\ \theta^*(B) = 1 - \theta_1^* B - \theta_2^* B^2 - \cdots - \theta_{q^*}^* B^{q^*} \end{cases} \tag{2}$$

Some of the parameters may be zero or otherwise constrained, so that (2) could be a multiplicative seasonal $ARIMA(p,d,q)(P,D,Q)_s$ model where

$$\begin{cases} \phi^*(B) = (1-B)^d (1-B^s)^D \phi(B)\Phi(B^s) \\ \theta^*(B) = \theta(B)\Theta(B^s) \end{cases} \tag{3}$$

Expressions $\phi(B)$, $\Phi(B^s)$, $\theta(B), \Theta(B^s)$ are polynomials of order, respec-

tively, p, P, q, Q and s indicates the length of the periodicity (seasonality). The same notation may be used to take into account multiple seasonality effects if necessary. Moreover, $p^* = p + s$, $q^* = q + sQ$. The notation $X_{t,j}, j = 1, 2, \cdots, m$ indicates m variables observed at day t influencing the price of electricity; β_j is a parameter measuring how the price P_t is related to the j-th variable $X_{t,j}$.

To keep the problem of estimating Equation (1) tractable, we use only deterministic exogenous variables so we know exactly what they will be at any future time (e.g. calendar variables, polynomials or sinusoids in time). The choice of known or non-stochastic regressors simplifies the inferential procedures, including estimation and testing of the parameters. This choice is also suggested by the fact that stochastic exogenous regressors, which must also be forecast, is one of the possible causes of inefficiency in prediction intervals. See [15] [Sect. 6.5]. According to the preparatory work for the present research, the most influential calendar variables are: days of the week, public holidays (official and religious) and daylight-saving time. Days immediately before and immediately after holidays are considered Saturdays and Mondays, respectively. Calendar effects are accounted for in the model by incorporating sets of dummy variables, where one of the categories is omitted to prevent complete collinearity. The dummy, or more precisely, binary variables in the process (1), precludes using the difference operators. It follows that, from now on the Box-Jenkins processes will be associated to the acronym SARMAX. The "burden of regular and seasonal non-stationarity" is placed entirely on the estimated parameters. In this sense, we require the roots of the polynomials $\phi^*(B)$ and $\theta^*(B)$ lie outside the unit circle, with no single root common to both polynomials. If this condition is satisfied then errors are stationary with finite variance.

The estimation of the parameters can be realized by optimizing the log-likelihood function of (1), provided that p, q, P, Q are known and errors are Gaussian random variables. Since we ignore the order of the polynomials, the estimation should be repeated for different values of p, q, P, Q. If

$$0 \le p \le n_p, 0 \le q \le n_q, 0 \le P \le n_P, 0 \le Q \le n_q,$$

then there are $(n_p + 1)(n_q + 1)(n_P + 1)(n_Q + 1)$ distinct processes to be explored for each time series. We execute the search of the best process in automatic mode, over a limited set of distinct variations by using the function *auto.arima*() of the *R* package forecast with the option of a stepwise search to reduce the high computational cost of brute force search.

A common index to compare rival models is the bias-corrected version of the Akaike criterion

$$\text{AICC} = -2n\left[\log\left(2\pi\hat{\sigma}_a^2\right) + \frac{\left(p^* + q^* + 1\right)}{n - p^* - q^* - 2}\right] \tag{4}$$

where $\hat{\sigma}_a^2$ denotes the estimated error variance of the candidate process. The process associated with the smallest AICC is presumed to be the best process. Let $L > 0$ be the number of prices to be foreseen (lead-time). The selected process

serves to compute, standing at time n, forecast $\hat{P}_{n+k}$ of the price at day $n+k, k=1,2,\cdots,L$ which are optimum in the sense of quadratic loss, conditional on an information set $I_n=\{P_1,P_2,\cdots,P_n\}$, *i.e.* $\hat{P}_{n+k}=E(P_{n+k}\mid I_n), k=1,2,\cdots,L$. It turns out that, under reasonably weak conditions, the optimal forecast is the expected value of the series being forecast, conditional on available information. See [16] (p. 172).

Forecasting the regression term in (1) does not present particular difficulties because of the perfectly predictable nature of the regressors. The future values of the stochastic process term can be computed by using the infinite moving-average representation of the optimal process

$$\left[\phi(B)\right]^{-1}\theta(B)a_t=\psi(B)a_t,\quad \text{with } \psi(B)=\sum_{i=0}^{\infty}\psi_i B^i, \psi_0=1. \tag{5}$$

where $\sum|\psi_i|<\infty$ (this constraint is equivalent to the requirement of roots outside the unit circle). The coefficients ψ_i in (5) are functions of the parameters in (2) and can be easily obtained by recursive equations. See [17]. In practice, however, the parameters of (2) have to be estimated, and it is customary to substitute estimated values into all the formulae.

4. Prediction Intervals

Short-term point forecasts cannot reflect all the uncertainties in the price of energy. In this regard, it is far more interesting to have information on how reliable the extent of the prediction is. In short, given a time series of n prices $P_1,P_2,\cdots,P_n$, we seek forecast limits such that the probability is $(1-\alpha)$ that P_{n+k} lies in

$$P_{n+k}\in\left[\hat{P}_{n,k}+Q_{n,k,(1-\alpha)/2},\hat{P}_{n,k}+Q_{n,k,(1+\alpha)/2}\right] \tag{6}$$

where $P_{n,k}$ is the price (/kWh) at a given hour of k days after day n and n is the last period at which a price is available. The point forecast $\hat{P}_{n,k}$ is obtained by identifying and estimating a SARMAX process to the fit period (*i.e.* training plus validation periods). $Q_{n,k,\alpha}$ is the quantile of order α of the distribution of the forecast error $e_{n,k}$ at origin n and lead-time k. If the hypothesis of Gaussianity is accepted for each k, then PIs can be derived from the standard formulae given by Box & Jenkins [18]

$$P_{n+k}\in\left[\hat{P}_{n,k}\pm z_{\alpha/2}\hat{\sigma}_k\right]. \tag{7}$$

where $z_{\alpha/2}$ is the upper $100(1-\alpha/2)\%$ point of the Gaussian distribution with zero mean and variance one. Moreover,

$$\hat{\sigma}_k^2=\hat{\sigma}_a^2\sum_{i=0}^{k-1}\hat{\psi}_i^2. \tag{8}$$

PIs in (7), typically called Box-Jenkins prediction intervals (BJ PIs), are the most commonly used even in the cases there are no specific reasons to assume a Gaussian distribution of the errors. [15] [Sect. 7.7] illustrates various possible reasons why PIs in (7) are inadequate to encompass the required proportion of

future prices and the lack of Gaussianity of the forecast errors is indicated as one of the causes. See also [19].

4.1. Williams Goodman Procedure

To simulate distribution of forecast errors, the time series is split into two parts: the "training" period and the "validation" period. As a preliminary, we choose a window size, ν (the number of consecutive daily prices) which, together with the maximum lead-time *L*, establish the complexity of the Williams Goodman (WG) procedure.

Initially, the training period contains prices for day 1 through ν whereas, the prices from $(\nu+1)$ to $(\nu+L)$ act as the first validation period. The class of SARMAX model discussed in Section 3 is fit to the training time series to find the best process which minimizes the AICC criterion (4). The selected process is then used to calculate the *L*-step-ahead point forecasts: $\hat{P}_{\nu+1,k}, k=1,2,\cdots,L$ at the time origin $\nu+1$. The post-sample forecast errors are obtained from difference with the corresponding values of the validation period:

$$\hat{e}_{\nu+1,k} = P_{\nu+1,k} - \hat{P}_{\nu+1,k}, k=1,2,\cdots,L.$$

Note that, in this case, $P_{\nu+1,k}$ is a real price and not a random quantity.

In the successive step, a block of γ contiguous prices is dropped from the start of the training period and, simultaneously, γ contiguous prices from the start of the validation period is shifted back to the end of the training period so that the second window contains prices for day $(1+\gamma)$ through day $(\nu+\gamma)$. The second validation period includes prices from $(\nu+\gamma+1)$ to $(\nu+\gamma+L)$ due to the inclusion of the next block of prices taken sequentially from the time points of the validation period not yet processed. The same class of models as in the initial step is fitted to the new training period, the new *L*-step-ahead forecasts calculated and the corresponding post-sample errors obtained at the time origin $\nu+\gamma+1$ as $\hat{e}_{h,\nu+\gamma+1,k} = P_{h,\nu+\gamma+1,k} - \hat{P}_{h,\nu+\gamma,k}, k=1,2,\cdots,L$.

The procedure is iterated until the last training period $(n-\nu-L+\gamma):(n-L)$ and the last validation period $(n-L+\gamma):n$ achieve the end *n* of the fit time series. Overall, the procedure forms $r=(n-\nu-L+1)$ distinct sequences of *L*-step-ahead forecast prices and post-sample forecast errors. We can arrange errors as a matrix.

$$\boldsymbol{G} = \begin{bmatrix} \hat{e}_{\nu+1,1} & \hat{e}_{h,\nu+1,2} & \cdots & \hat{e}_{h,\nu+1,L} \\ \hat{e}_{\nu+2,1} & \hat{e}_{\nu+2,2} & \cdots & \hat{e}_{\nu+2,L} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{e}_{\nu+r,1} & \hat{e}_{\nu+r,2} & \cdots & \hat{e}_{\nu+r,L} \end{bmatrix} \tag{9}$$

Rows correspond to different time origins and columns to different lead-times. If the forecast error distributions are the same, then column $\boldsymbol{g}_k$ can be intended as a sample of size *r* of the forecast errors that would have been made by the selected SARMAX process, at lead-time *k* across horizon origins $\nu+1, \nu+2, \cdots, \nu+r$.

The construction of PIs requires knowledge of the quantiles of the forecast error distribution, which are typically unknown and have to be estimated. An obvious way to generate PIs is to assume *k*-step-ahead forecast errors follow a continuous distribution function. If the fitting is satisfactory, the quantiles of the distribution can be estimated and then prediction bounds determined for each lead-time.

Chatfield [15] [Sect. 7.5.5], notice that, although the WG procedure is attractive in principle, it seems to have been underutilized, not only because of the lack of time series long enough to give credibility to the fit of empirical distributions, but also because of the heavy computational requirements involved. Of course, the length of the time series is not a problem with time series collected in the electricity market and analyzed in the present study. In addition, the effort required to implement WG for time series of moderate length (1000 - 1500 time points) is compatible with the hardware/software resources generally at disposal. An R script is available from the authors upon request.

4.2. Density Selection

In the framework of electricity price forecasting, it might be reasonably argued that prices are not Gaussian (see, [7] [10]) but it is not clear what is precisely their distribution. In this subsection, however, we discuss the distributional properties of post-sample forecast errors whose behavior cannot be automatically deduced with certainty from observed values and/or in-sample forecast errors. If the empirical situation does not suggest an obvious choice, one can be selected among myriad examples of probability density functions (pdfs). Williams & Goodman [1] adapted a gamma density to the absolute value of the forecast errors. Isengildina-Massa *et al.* [20] found that forecast errors from the data set used in their study most often followed a logistic distribution. Bordignon & Lisi [21] and Lee & Scholtes [22] propose the Gaussian distribution. We choose the mathematical function in the first row of **Table 1** after a long and complicated search for a powerful and versatile curve. Obviously, we are aware that, trying many densities and keeping the "best fitting" one, does not guarantee that another model will not look better than those we have already seen.

In all the densities, θ_1 controls the location of the distribution; $\theta_2 > 0$ affects the scale, θ_3 and θ_4 are shape parameters. The densities are referred to

Table 1. Density functions for post-sample forecast errors.

Name	Support	Model	*R* package	Est.Met.
Johnson's system	$(-\infty,\infty)$	$\frac{\theta_3}{\sqrt{2\pi}} g'(y) \mathrm{e}^{[\theta_4+\theta_3 g(y)]^{-2}/2}$	SuppDists	qme
Gamma	$(0,\infty)$	$\frac{y^{\theta_3-1}\mathrm{e}^{-y}}{\theta_2\Gamma(\theta_3)}$	fitdistrplus	mme
Logistic	$(-\infty,\infty)$	$\frac{\mathrm{e}^{-y}}{(1+\mathrm{e}^{-y})^2}$	stats	mme
Gaussian	$(-\infty,\infty)$	$\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-y^2/2}$	stats	mle

$y=(x-\theta_1)/\theta_2$. The gamma density is fit to the absolute value of post-sample errors and hence $\theta_1=0$. The system proposed by Johnson [23] contains three classes of distributions which are based on the transformation $z=\theta_4+\theta_3\log[g(y)]$ where z is a standard Gaussian random variable and g (with derivative g') has three possible forms

$$\begin{aligned} &S_L: g(y)=y && \text{the logGaussian} \\ &S_U: g(y)=y+\sqrt{1+y^2} && \text{an unbounded distribution} \\ &S_B: g(y)=y/(1-y) && \text{a bounded distribution} \end{aligned} \tag{10}$$

In using (10), a first problem to be solved is to determine which of the three families should be used and, once the class is selected, the next problem to be solved is to estimate the parameters. In both problems, we follow the technique proposed by Wheeler [24] as implemented in package in the package *SuppDists* with quantiles $(0.1, 0.25, 0.5, 0.75, 0.9)$.

The column headed "*R* package" refers to the package used for parameter estimation. The notation "stats" indicates standard computational algorithms. The last column of **Table 1** reports the technique of estimation of the parameters: mle (maximum likelihood), qme (quantile method), mme (method of moments).

The usual strategy behind fitting a given distribution to data is to identify the type of density curve and estimate the parameters that give the highest probability of producing the observed values. Instead, we follow an indirect approach: we compare the different density curves by testing how accurately the PIs generated by a SARMAX process, in tandem with Williams-Goodman method, capture the true prices.

5. Forecasting Accuracy

Let us consider the matrix of estimated forecast errors $\boldsymbol{G}$ discussed in the preceding section. For each lead-time $k=1,2,\cdots,L$ we fit the distributions shown in **Table 1**. Let $\hat{Q}_{k,v,\alpha}$ be the α-th estimated quantile of the v-th distribution, $v=1,2,3$. The quantiles are used to derive $(1-\alpha)$ parametric prediction limits

$$C_{k,v}^1: \hat{P}_{n,k}+\tilde{\mu}_{n,k}+\hat{Q}_{k,v,(1-\alpha)/2}\tilde{\sigma}_{n,k}, \quad C_{k,v}^2: \hat{P}_{n,k}+\tilde{\mu}_{n,k}+\hat{Q}_{k,v,(1+\alpha)/2}\tilde{\sigma}_{n,k} \tag{11}$$

The means and standard deviations are computed over the post-sample errors

$$\tilde{\mu}_{n,k}=\frac{1}{v}\sum_{t=1}^{v}\hat{e}_{v+t,k}, \quad \tilde{\sigma}_{n,k}^2=\frac{1}{v}\sum_{t=1}^{v}\left[\hat{e}_{v+t,k}-\tilde{\mu}_{n,k}\right]^2, \quad k=1,2,\cdots,L \tag{12}$$

Notice that the mean of post-sample errors $\tilde{\mu}_{n,k}$ is not necessarily zero.

To assess the performance of the various PIs, we compare the prediction interval actual coverage (*PIAC*) to $(1-\alpha)\%$. The *PIAC* is measured by counting the number of true hourly prices of the target period enclosed in the bounds (11)

$$PIAC_v=100L^{-1}\sum_{k=1}^{L}c_{k,v} \quad \text{where } c_{k,v}=\begin{cases}1 & \text{if } P_{n+k}\in\left[C_{k,v}^1, C_{k,v}^2\right] \\ 0 & \text{otherwise}\end{cases} \tag{13}$$

If the PIs are accurate, then $PIAC_v \geq (1-\alpha)$. All other things being equal, nar-

row PIs are desirable as they reduce the uncertainty associated with forecast-based decision making. However, there is a trade-off between PI widths and *PIAC*: the wider the PI, the higher the corresponding *PIAC* and hence the greater is the accuracy of predictions, at least to a certain extent, because very wide PIs are not practically useful. On the other hand, very sharp PIs with a low coverage probability are useless as well. It is necessary to introduce, in this connection, a scoring rule for addressing the sharpness of PIs. We use a score function of the form proposed by Winkler & Murphy [25]

$$S_{v,k} = \left(\frac{1-\alpha}{2}\right)\frac{\left(C_{k,v}^{2} - C_{k,v}^{1}\right)}{\left|P_{n+k}\right|} + I\left(P_{n+k} < C_{k,v}^{1}\right)\frac{\left(C_{k,v}^{1} - P_{n+k}\right)}{\left|P_{n+k}\right|} + I\left(P_{n+k} > C_{k,v}^{2}\right)\frac{\left(P_{n+k} - C_{k,v}^{2}\right)}{\left|P_{n+k}\right|},\ k = 1, 2, \cdots, L \tag{14}$$

The use of ratios facilitates comparability across price levels. The symbol $I(\,)$ represents an indicator function taking one if the argument is true and zero otherwise. The first addend in (14) reflects a cost associated with the width of the interval. The cost decreases as $(1-\alpha)$ increases, to compensate the tendency of the bounds to be broader as the confidence level increases. The other two addends penalize PIs if the target is outside the interval. The penalty increases with increasing distance from the nearest interval endpoints. The average of (14) across time points provides an indication of the sharpness of PIs

$$MS_{v} = \frac{1}{L}\sum_{k=1}^{L} S_{v,k} \tag{15}$$

Criteria (13) and (15) should be judged keeping in mind the stochastic behavior of the electricity prices. Here, we have a potentially severe problem. Price peaks and valleys have been smoothed for the training and validation periods, but the same has not been done for the target period. These prices, in fact, are left as they are observed, to simulate real conditions. Spike prices, however, are recurring events and, therefore, it would not be surprising to find some of them in the target time series. Our SARMAX processes, being developed within a cleaned-up data set, have hardly any possibility to predict satisfactorily all, or at least a good part, of the outliers. Remaining outliers imply poor prediction intervals in practice. Further research is required to formulate a model which is not only generally enough to merge Box-Jenkins processes, WG prediction intervals and spike prices, but also it is numerically tractable to provide a quantitative description of the complex patterns of electricity market time series.

Predictive Performance

To this end, we analyze $144 = 24 \times 6$ different time series, one for each hour of the day and each zone of the Italian electricity market. All the daily time series are long 1148 days, but the last three weeks ($L = 21$) are reserved for assessing the predictive accuracy of the intervals. Thus, only the first 1127 days are used for estimation and validation of SARMAX models. The size of the rolling win-

dow is fixed at $v = 959$ (15%) which leads to $r = 168$ samples of 21-step ahead forecasts. The search of the SARMAX processes is conducted within the bounds $n_p = n_q = n_P = n_Q = 2$, which include 81 different processes. Each process is combined with the WG procedure applied to any one of the density functions in Table 1. For completeness, we include the BJ PIs described in (7) and Tchebycheff PIs

$$Pr\left[\hat{P}_{n,k} + \tilde{\mu}_{n,k} - c\tilde{\sigma}_k < P_{n+k} < \hat{P}_{n,k} + \tilde{\mu}_{n,k} + c\tilde{\sigma}_k\right] \geq 1 - \frac{1}{c^2}, \text{ with } c = (1-\alpha)^{-0.5} \quad (16)$$

At this point it is necessary to remind that the point estimates are obtained for log-prices $P_{n,k}$, which have to be transformed back to the original scale to give forecasts for $e^{P_{n,k}}$. A simply way is to transform the bounds obtained for $P_{n,k}$ by applying a simple fractional multiplier

$$\delta e^{\hat{P}_{n,k} + \tilde{\mu}_{n,k} + \hat{Q}_{k,v,(1-\alpha)/2}\tilde{\sigma}_{n,k}} < e^{P_{n+k}} < \delta e^{\hat{P}_{n,k} + \tilde{\mu}_{n,k} + c\tilde{\sigma}_k < P_{n+k}} \quad (17)$$

where $\delta = e^{0.5\hat{\sigma}_k^2}$ is the correction factor proposed by [Baskerville, 1972] and $\hat{\sigma}_k$ is the standard deviation introduced in (8). Expression (17) is a genuine $(1-\alpha)$ PI since $e^{P_{n,k}}$ is a monotone function of P_{n+k}, but they do not have necessarily the same structure. For example, the interval (17) is asymmetrical even though the interval in the log scale is symmetrical. Furthermore, the anti-logarithms of forecasts are biased. See [26].

Table 2 shows the results of *PIAC* and *MS* at $(60, 65, 70, \cdots, 90, 95)$. The row

Table 2. Quality of PiIs.

$(1-\alpha)$%		Johnson's	Gamma	Logistic	Gaussian	BJ	Tchebycheff
60	*PIAC*	75.7	77.6	68.0	65.5	83.3	81.9
	MS	5.9	9.3	6.6	7.6	53.1	111.0
	Frct	36.1	14.6	24.3	8.3	9.7	6.9
65	*PIAC*	78.4	77.6	73.6	69.4	85.7	81.7
	MS	5.6	9.8	5.9	7.7	51.7	169.0
	Frct	43.8	14.6	19.4	8.3	9.7	4.2
70	*PIAC*	82.7	83.6	76.5	73.5	87.1	-
	MS	5.0	8.2	5.7	6.6	50.2	-
	Frct	46.5	12.5	18.8	4.9	9.7	-
75	*PIAC*	85.7	84.2	83.5	78.8	90.9	87.4
	MS	4.5	8.0	5.0	5.6	37.2	110.0
	Frct	49.3	9.0	16.7	9.0	8.3	7.6
80	*PIAC*	88.5	89.9	84.6	86.5	94.4	92.9
	MS	3.9	6.4	4.6	4.9	37.8	84.2
	Frct	55.6	5.6	17.4	4.2	7.6	9.7
85	*PIAC*	91.3	92.9	92.3	88.9	95.7	94.6
	MS	3.1	5.8	3.7	4.1	4.1	90.3

Continued

	Frct	54.9	5.6	16.7	4.2	7.6	11.1
90	*PIAC*	94.8	95.2	96.4	90.5	97.4	97.9
	MS	2.3	4.3	2.9	2.0	3.1	80.9
	Frct	56.2	5.6	20.1	0.7	6.2	11.1
95	*PIAC*	97.7	95.2	-	98.3	97.1	97.0
	MS	1.4	2.5	-	1.7	1.8	91.9
	Frct	52.1	4.9	-	28.5	6.9	7.6

entitled "Frct" denotes the percentage of times, out of the 144 cases studied, the corresponding density determined PIs with the lowest magnitude among all the PIs associated with an actual coverage rate greater than or equal to $(1-\alpha)$. The symbol "-" indicates that the corresponding distribution has never taken the first place in the two rankings of forecast accuracy used in this study.

On a first general examination, we note the consistent behavior between actual coverage rate (*PIAC*) and mean relative scores (*MS*), with the latter decreasing as the former increases. Naturally, this is a confirmation of the expected behavior of the score function (14). Tchebycheff intervals show, perhaps not surprisingly, the largest widths. Box-Jenkins prediction intervals (BJ PIs) appear to be the most conservative approach, *i.e.* it yields largest coverage rates, but with prevalently smaller widths than the Tchebycheff PIs. The substantial reliability in performance of the WG procedure based on Johnson's system, gamma, logistic and Gaussian distributions is due to actual coverage rates which are decidedly lower than the corresponding nominal coverage rates than BJ and Tchebycheff PIs. But, above all, the formers have a much smaller sharpness than the latters at all confidence levels. The distributions nested within Johnson's system come more frequently on top in terms of actual coverage probability and sharpness of the intervals and hence can be considered the optimal probability density within the experimental set up of our study.

6. Conclusions

Prediction intervals (PIs) are random sets designed to contain a future value with a given probability. The principal reason for constructing them is to provide an indication of the reliability of point forecasts avoiding a complete description of the probability distribution of the uncertainty associated with a prediction. Box-Jenkins or BJ PIs (the procedure in common use currently) assume Gaussian errors, known parameters and intervals are centered about the conditional expectation. Consequently, BJ PIs cannot take into account the variability due to parameter estimation and behave poorly when the errors are not Gaussian. Our findings confirm these observations.

The primary concern in this paper is with the Williams & Goodman [1] (WG) procedure and the gain of accuracy brought by WG PIs in comparison to con-

ventional BJ PIs. Our findings show point forecasts for day-ahead hourly prices in Italian wholesale electricity market may be of greater utility if accompanied by PIs calculated using WG centered around a member of the Johnson system of density functions. This mix offers PIs having a coverage rate greater than the nominal rate, but, what is more important is that this desirable conservativeness is not at the expense of widening the intervals, which instead are admirably sharp. The procedure has proven to be very competitive on two other procedures: Box-Jenkins and Tchebycheff PIs.

Acknowledgements

The authors would like to thank a referee for helpful comments, which lead to an improvement of the paper.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Williams, W.H. and Goodman, M.L. (1971) A Simple Method for the Construction of Empirical Confidence Limits for Economic Forecasts. *Journal of the American Statistical Association*, **66**, 752-754. https://doi.org/10.1080/01621459.1971.10482340

[2] Misiorek, A., Trueck, S. and Weron, R. (2006) Point and Interval Forecasting of Spot Electricity Prices: Linear vs. Non-Linear Time Series Models. *Studies in Nonlinear Dynamics and Econometrics*, **10**. Article 2. https://doi.org/10.2202/1558-3708.1362

[3] Weron, R. (2006) Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach. John Wiley & Sons, Chichester. https://doi.org/10.1002/9781118673362

[4] Wu, H.C., Chan, S.C., Tsui, K.M. and Hou, Y. (2013) A New Recursive Dynamic Factor Analysis for Point and Interval Forecast of Electricity Price. *IEEE Transactions on Power Systems*, **28**, 2352-2365. https://doi.org/10.1109/TPWRS.2012.2232314

[5] Nowotarski, J. and Weron, R. (2015) Computing Electricity Spot Price Prediction Intervals Using Quantile Regression and Forecast Averaging. *Computational Statistics*, **30**, 791-803. https://doi.org/10.1007/s00180-014-0523-0

[6] Bunn, D., Andresen, A., Chen, D. and Westgaard, S. (2016) Analysis and Forecasting of Electricity Price Risks with Quantile Factor Models. *The Energy Journal*, **37**, 101-122. https://doi.org/10.5547/01956574.37.1.dbun

[7] Nowotarski, J. and Weron, R. (2017) Recent Advances in Electricity Price Forecasting: A Review of Probabilistic Forecasting. *Renewable & Sustainable Energy Reviews*, in press.

[8] Borovkova, S. and Schmeck, M.D. (2017) Electricity Price Modeling with Stochastic Time Change. *Energy Economics*, **63**, 51-65. https://doi.org/10.1016/j.eneco.2017.01.002

[9] Gianfreda, A. and Grossi, L. (2012) Forecasting Italian Electricity Zonal Prices with

Exogenous Variables. *Energy Economics*, **34**, 2228-2239. https://doi.org/10.1016/j.eneco.2012.06.024

[10] Knittel, K.R. and Roberts, M.R. (2005) An Empirical Examination of Restructured Electricity Prices. *Energy Economics*, **27**, 791-817. https://doi.org/10.1016/j.eneco.2004.11.005

[11] Fildes, R. (1992) The Evaluation of Extrapolative Forecasting Methods. *International Journal of Forecasting*, **8**, 81-98. https://doi.org/10.1016/0169-2070(92)90009-X

[12] Conejo, A.J., Contreras, J., Espínola, R. and Plazas, M.A. (2005) Forecasting Electricity Prices for a Day-Ahead Pool-Based Electric Energy Market. *International Journal of Forecasting*, **21**, 435-462. https://doi.org/10.1016/j.ijforecast.2004.12.005

[13] Voronin, S. and Partanen, J. (2013) Price Forecasting in the Day-Ahead Energy Market by an Iterative Method with Separate Normal Price and Price Spike Frameworks. *Energies*, **6**, 5897-5920. https://doi.org/10.3390/en6115897

[14] Weron, R. (2014) Electricity Price Forecasting: A Review of the State-of-the-Art with a Look into the Future. *International Journal of Forecasting*, **30**, 1030-1081. https://doi.org/10.1016/j.ijforecast.2014.08.008

[15] Chatfield, C. (2000) Time Series Forecasting. Chapman & Hall/CRC, Boca Raton. https://doi.org/10.1201/9781420036206

[16] Diebold, F.X. (2007) Elements of Forecasting. 4th Edition, South-Western, Mason. http://threeplusone.com/fieldguide

[17] Cheung, S.H., Wu, K.H. and Chart, W.S. (1998) Simultaneous Prediction Intervals for Autoregressive-Integrated Moving-Average Models: A Comparative Study. *Computational Statistics* & *Data Analysis*, **28**, 297-306. https://doi.org/10.1016/S0167-9473(98)00038-3

[18] Box, G.E.P. and Jenkins, G.M. (1976) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

[19] Taylor, J.W. and Bunn, D.W. (1999) Investigating Improvements in the Accuracy of Prediction Intervals for Combinations of Forecasts: A Simulation Study. *International Journal of Forecasting*, **15**, 325-339. https://doi.org/10.1016/S0169-2070(99)00002-3

[20] Isengildina-Massa, O., Irwin, S., Good, D.L. and Massa, L. (2011) Empirical Confidence Intervals for USDA Commodity Price Forecasts. *Applied Economics*, **43**, 3789-3803. https://doi.org/10.1080/00036841003724429

[21] Bordignon, S. and Lisi, F. (2001) Interval Prediction for Chaotic Time Series. *Metron*, **59**, 117-140.

[22] Lee, Y.S. and Scholtes, S. (2014) Empirical Prediction Intervals Revisited. *International Journal of Forecasting*, **30**, 217-234. https://doi.org/10.1016/j.ijforecast.2013.07.018

[23] Johnson, N.L. (1949) Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*, **36**, 149-176. https://doi.org/10.1093/biomet/36.1-2.149

[24] Wheeler, R.E. (1980) Quantile Estimators of Johnson Curve Parameters. *Biometrika*, **67**, 725-728. https://doi.org/10.1093/biomet/67.3.725

[25] Winkler, R.L. and Murphy, A.H. (1979) The Use of Probabilities in Forecasts of Maximum and Minimum Temperatures. *The Meteorological Magazine*, **108**, 317-329.

[26] Clifford, D. and Cressie, N. and England, J.R. and Roxburgh, S.H. and Paul, K.I.

(2013) Correction Factors for Unbiased, Efficient Estimation and Prediction of Biomass from Log-Log Allometric Models. *Forest Ecology and Management*, **310**, 375-381. https://doi.org/10.1016/j.foreco.2013.08.041