CS-BIGS 3(2): 79-85
http://www.bentley.edu/csbigs/blanke.pdf

Modeling and Forecasting CPC Prices

Delphine Blanke
Université d'Avignon et des Pays de Vaucluse, France

Denis Bosq
Université Pierre et Marie Curie, Paris, France

This paper deals with modeling and forecasting coke petroleum calcination (CPC) prices. We first consider some empirical techniques (polynomial regression, Holt and Winters smoothing) and compare them with the more general Box and Jenkins method. We also use nonparametric predictors. The case study is accessible to readers with an intermediate level of statistics. Prior exposure to Box-Jenkins techniques is useful but not strictly necessary.

1. Introduction

The origin of the current study was a request concerning a privatization project of a company, requiring its valuation. Current market conditions for CPC are involved in the future turnover of this project. CPC (Coke Petroleum Calcination) is derived from a by-product of oil and is used in aluminum and titanium alloys. In this paper, we study the structure and prediction of CPC prices (in dollars) from quarterly data (denoted from Q1 to Q4) between 1985-Q1 and 2008-Q4 (figure 1). Due to the irregular variability of these data, the problem is somewhat intricate.

In a first report, an expert used simple linear regression and obtained not very plausible results. The goal of this paper is to compare linear regression (LR) with more efficient methods: parabolic regression (PR), Holt and Winters filtering (HW), the Box and Jenkins method (BJ) and finally, nonparametric prediction (NP). For this purpose, we have used the software *R* developed by the R Development Core Team (2008).

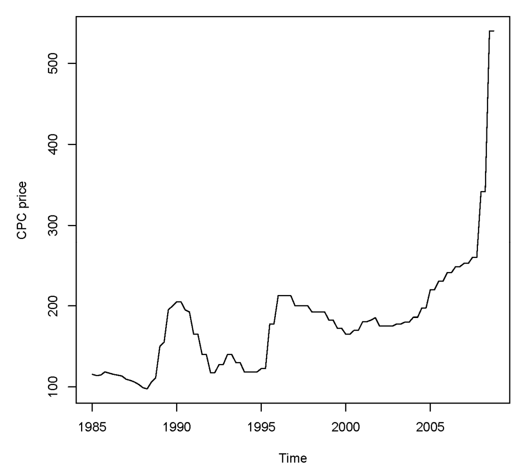

Figure 1: CPC prices from 1985-Q1 to 2008-Q4

It is noteworthy that LR, PR and HW may be considered as special cases of BJ (see Section 2). Thus, not surprisingly, BJ appears as more efficient than the former techniques.

The non-homogeneity of the data led us to divide them into three parts: 1985-Q1 to 1995-Q4 where a seasonal component appears, 1996-Q1 to 2007-Q4 where the trend is parabolic and 2008 data that can be considered as outliers (possibly due to the economic crisis). In this paper, our aim is to construct forecasts, especially for 2009, not taking into account the 2008 exotic data (since they are not representative of typical conditions upon which to build an evaluation of the company).

The next section deals with polynomial regression and HW. We specify the link of these methods with BJ and explain why LR and PR are not suitable for CPC study. The third part is devoted to BJ. We obtain two different models: one for 1985-Q1 to 1995-Q4, another one for 1996-Q1 to 2007-Q4. Finally, the NP method is considered in Section 4.

2. Empirical Methods

The **linear regression** (LR) model has the form:

$$X_t = a_0 + a_1 t + \varepsilon_t, \quad t \in \mathbb{Z}$$

where (X_t) is the observed process, α_0 and α_1 are real coefficients and (ε_t) is a white noise:

$$E\varepsilon_t^2 = \sigma^2 > 0, E\varepsilon_t = 0,$$
$$E(\varepsilon_s \varepsilon_t) = 0\,;\, s,t \in \mathbb{Z}, s \neq t.$$

We first show that, in some sense, the LR model is a special ARIMA (Auto Regressive Integrated Moving Average) process (ARIMA theory appears in Brockwell and Davis (1991) among other references). Set

$$Y_t = X_t - X_{t-1} - a_1, \quad t \in \mathbb{Z} \tag{1}$$

then

$$Y_t = \varepsilon_t - \varepsilon_{t-1}, \quad t \in \mathbb{Z} \tag{2}$$

and (Y_t) is a MA(1), hence (X_t) is a non-centered ARIMA(0,1,1).

In addition, despite the fact that the polynomial of degree one which appears in (2) has a unit root, (ε_t) is the innovation process of (Y_t). In order to prove that assertion, we first note that (2) implies

$$\varepsilon_t = Y_t + \cdots + Y_{t-j} + \varepsilon_{t-j-1}, \quad j \geq 0. \tag{3}$$

Then, using (3) for $j = 0, \cdots, k-1$, one obtains

$$\varepsilon_t = \sum_{j=0}^{k-1} (1 - \frac{j}{k}) Y_{t-j} + \frac{1}{k} \sum_{j=0}^{k-1} \varepsilon_{t-j-1}$$

and since, for t fixed,

$$E(\frac{1}{k}\sum_{j=0}^{k-1} \varepsilon_{t-j-1})^2 = \frac{\sigma^2}{k} \xrightarrow[k\to\infty]{} 0,$$

it follows that

$$\sum_{j=0}^{k-1} (1 - \frac{j}{k}) Y_{t-j} \xrightarrow[k\to\infty]{L^2} \varepsilon_t$$

where $\xrightarrow{L^2}$ stands for convergence in mean square. Consequently, $\varepsilon_t \in M_t$, the closed linear space generated by $Y_s, s \leq t$. Noting that ε_t is orthogonal to M_{t-1}, we conclude that (ε_t) is the innovation of (Y_t) and that $-\varepsilon_{t-1}$ is the orthogonal projection of Y_t on M_{t-1}, that is the best linear predictor of Y_t given $Y_s, s \leq t-1$.

Now the LR is not convenient for data with irregular variations like CPC since it only computes a trend, assuming that this trend is linear, and does not take into account the correlation between the X_t's. This limitation appears in Figure 2 and 3 where the "explained variances" R^2 are respectively 7% and 28% !

The **parabolic regression** (PR) model is written as

$$X_t = a_0 + a_1 t + a_2 t^2 + \varepsilon_t, \quad t \in \mathbb{Z}. \tag{4}$$

In order to give an ARIMA interpretation of PR, we differentiate to obtain

$$X_t - X_{t-1} = a_2(2t-1) + a_1 + \varepsilon_t - \varepsilon_{t-1}$$

a second differentiation leads to the relation

$$X_t - 2X_{t-1} + X_{t-2} = 2a_2 + \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2}.$$

If we put

$$Y_t = X_t - 2X_{t-1} + X_{t-2} - 2a_2$$

it follows that (Y_t) is a MA(2) and (X_t) becomes a non-centered ARIMA(0,2,2).

Again the polynomial associated with (Y_t) has a (double) unit root and (ε_t) is the innovation of (Y_t). In order to prove that claim, we set

$$E_t = \varepsilon_t - \varepsilon_{t-1}, \quad t \in \mathbb{Z}$$

then we have

$$Y_t = E_t - E_{t-1}$$

and, using a similar method as above, we obtain the relation $E_t \in M_t$ and

$$\varepsilon_t = \sum_{j=0}^{k-1} (1 - \frac{j}{k}) E_{t-j} + \frac{1}{k} \sum_{j=0}^{k-1} \varepsilon_{t-j-1}.$$

Letting k tending to infinity gives $\varepsilon_t \in M_t$ and $\varepsilon_t \perp M_{t-1}$, the proof is therefore complete.

PR suffers for the same limitation as LR but fits CCP prices from 1996 to 2007 better with a R^2 of 93% (Figure

3). Note however that the R^2 is not a completely satisfactory criterion of efficiency, we refer to Mélard (1990) for a comprehensive discussion.

The **Holt and Winters** method is more sophisticated because it also takes into account a possible seasonality of the data. For an additive seasonal model with period length p and if $\hat{X}_T(H)$ denotes the prediction of X_{T+H} given the data $X_1,\ldots X_T$, one has the additive Holt-Winters forecast :

$$\hat{X}_T(H) = a_T + Hb_T + S_{T+1+(H-1) \bmod p}$$

where a_T , b_T and S_T are recursively given by

$$a_T = \alpha\ (X_T - S_{T-p}) + (1-\alpha)(a_{T-1} + b_{T-1})$$

$$b_T = \beta(a_T - a_{T-1}) + (1-\beta)b_{T-1}$$

$$S_T = \gamma(X_T - a_T) + (1-\gamma)S_{T-p}$$

and where the smoothing parameters α , β and γ are taken in [0,1]. Their optimal values are determined by minimizing the squared one-step prediction error. The functions a , b and S are initialized by performing a simple decomposition in trend and seasonal component and using moving averages on the first periods. This method is optimal for a SARIMA$(0,2,2)(0,1,1)^P$ model (here the triplet (0,2,2) refers to the AR, differencing and MA orders for the series, the triplet (0,1,1) to the same orders for the seasonal components and P to the length of the seasonal period, see, for example, Bosq and Lecoutre, 1992). The results obtained by this method are quite satisfactory for both periods (see Figure 2 and 3).

3. The BJ Method

We have seen that the previous techniques are associated with various SARIMA models. Then it is natural to use the BJ method for modeling and forecasting our data.

For the period 1985-1995 we obtain a SARIMA (2,0,2) $(1,0,1)^4$: the seasonality period is one year (four quarters) and there is no trend (see Figure 2). However, the forecasts for 1997 and 1998 are not completely satisfactory, indicating a probable change of model. For 1996-2007 the new model is ARIMA (2,2,1) (see Figure 3). Thus, the seasonality has disappeared and the trend is parabolic (I=2).

The forecasts and the prediction intervals (at confidence level 0.95) are presented in Figure 4. Recall that our main interest is to predict 2009 postulating a return to a quieter situation. Results obtained by LR are not included in these prediction intervals but PR appears as an upper limit of them. In addition, we note that the model remains the same by considering only the data from 1996-

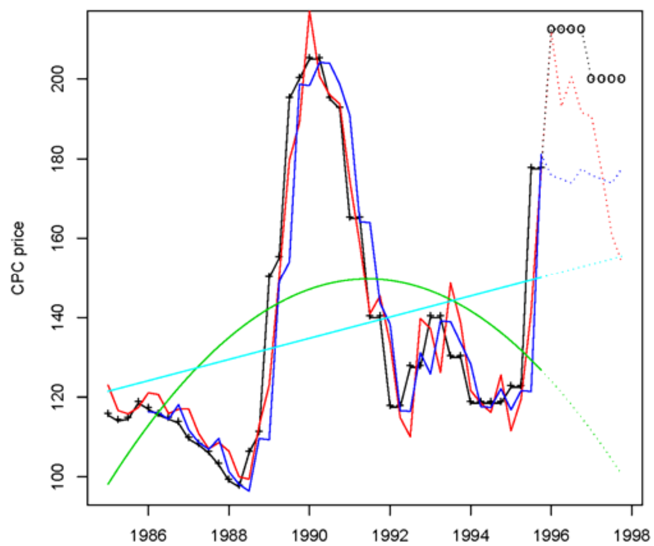

Figure 2: CPC predictions built with data of 1985-Q1 to 1995-Q4

Legend:

Black line and crosses: observed CPC prices (1985-I to 1995-IV)
Black circles: observed CPC prices (1996-Q1 to 1997-Q4)
Red: modelized values by BJ
Blue: modelized values by HW
Green: modelized values by PR
Cyan: modelized values by LR
Dotted: predictions by BJ, HW, PR and LR from 1996-Q1 to 1997-Q4

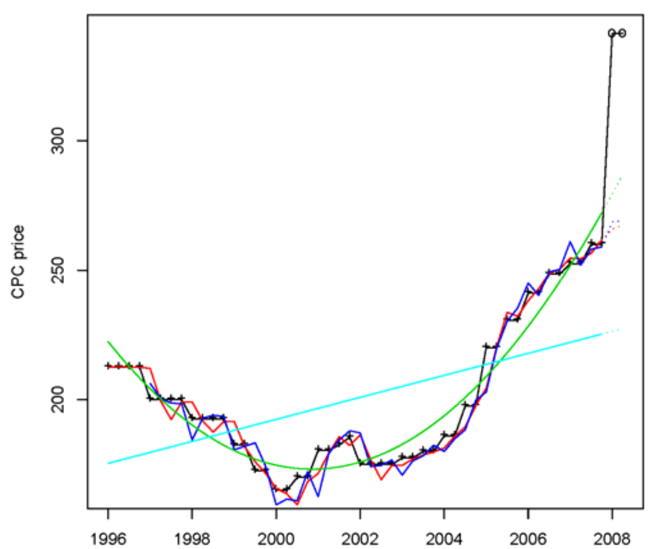

Figure 3: CPC predictions built with data of 1996-Q1 to 2007-Q4

Legend:

Black line and crosses: observed CPC prices (1996-Q1 to 2007-Q4)
Black circles: observed CPC prices (2008-Q1 to 2008-Q2)
Red (resp. blue, green and cyan): modelized values by BJ (resp by HW, by PR and LR)
Dotted: predictions by BJ, HW, PR and LR for 2008-Q1 and 2008-Q2

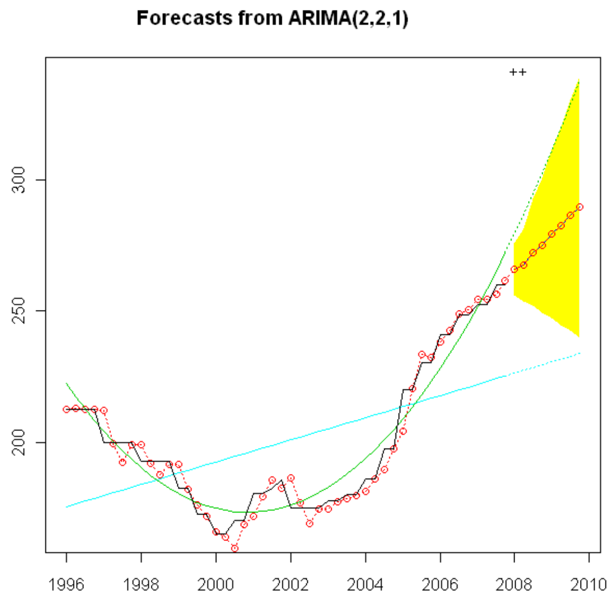

Figure 4: Prediction intervals for CPC by BJ method
Legend:
Black line: observed CPC prices (1996-Q1 to 2007-Q4)
Black crosses: 2008-Q1 and 2008-Q2 observed CPC prices
In red: modelized values by BJ
In green: modelized values by PR
In cyan: modelized values by LR
Dotted: predictions by PR and LR for 2008-Q1 to 2009-Q4
Yellow zone : prediction interval by BJ for 2008-Q1 to 2009-Q4 at confidence level 0.95

Q1 to 2006-Q4 leading to sharp forecasts of the observed year 2007. On the contrary, if one adds the two first data points of 2008, the model becomes an ARIMA (1,2,0) but with a lesser fit.

4. The Nonparametric Method

The BJ method postulates that the underlying model is of SARIMA type. That assumption being somewhat arbitrary it is often convenient to employ a nonparametric method (avoiding the estimation of a possibly important number of parameters). This method is, in some sense, "objective" since the underlying model only appears through regularity conditions.

If Y is a stationary and Markovian process observed at instants $1,\ldots,n$, the nonparametric predictor of Y_{n+H} (where $H \geq 1$ represents the horizon) takes the form:

$$\hat{Y}_{n+H} = \hat{r}_{n-H}(Y_n)$$

where $\hat{r}_{n-H}(x)$ is a nonparametric estimator of the regression $E(Y_H / Y_0 = x)$ based on the data (Y_{i+H}, Y_i) for $i = 1,\ldots,n-H$. For example, one may choose the popular kernel method with the Nadaraya-Watson (NW) estimator. Construction of that predictor, at horizon $H \geq 1$, may be described as follows: suppose that (Y_i, Y_{i+H}) has a density $f(x,y)$ which does not depend on i , then

$$E(Y_{i+H} / Y_i = x) = \int_{-\infty}^{\infty} y f(x,y)\mathrm{d}y \Big/ \int_{-\infty}^{\infty} f(x,y)\mathrm{d}y. \quad (5)$$

Now a simple histogram type estimator of f is

$$f_n^0(x,y) = \frac{1}{(n-H)h_n^2} \sum_{i=1}^{n-H} 1_{[x-\frac{h_n}{2}, x+\frac{h_n}{2}]}(Y_i) 1_{[y-\frac{h_n}{2}, y+\frac{h_n}{2}]}(Y_{i+H})$$

with bandwidth h_n such that $\lim_{n\to\infty} h_n = 0$. A smooth version of f_n^0 has the form

$$f_n(x,y) = \frac{1}{(n-H)h_n^2} \sum_{i=1}^{n-H} K\left(\frac{x-Y_i}{h_n}\right) K\left(\frac{y-Y_{i+H}}{h_n}\right)$$

where K is a strictly positive symmetric continuous probability density (f_n^0 corresponding to the choice $K = 1_{[-\frac{1}{2},\frac{1}{2}]}$). Replacing f by f_n in (4) leads to

$$\hat{r}_{n-H}(x) = \sum_{i=1}^{n-H} Y_{i+H} K\left(\frac{x-Y_i}{h_n}\right) \Big/ \sum_{i=1}^{n-H} K\left(\frac{x-Y_i}{h_n}\right)$$

then, if (Y_n) is Markovian, the NW predictor of Y_{n+H} is

$$\hat{Y}_{n+H} = \sum_{i=1}^{n-H} Y_{i+H} K\left(\frac{Y_n-Y_i}{h_n}\right) \Big/ \sum_{i=1}^{n-H} K\left(\frac{Y_n-Y_i}{h_n}\right).$$

Prediction results for stochastic processes by kernel appear in Bosq and Blanke (2007).

Based on the BJ model, we construct the nonparametric predictors on the twice differentiated data from 1996-Q1 to 2007-Q4 which can be considered as a stationary process. The results are detailed in Annex A5.

5. Conclusion

We have seen that the LR, PR and HW methods are all associated with SARIMA models. Since the BJ method selects the best SARIMA model, it is natural to consider the BJ forecasts as "optimal". Finally, it is interesting to note that the nonparametric predictions are close to the BJ ones.

Acknowledgments

We are grateful to S. Abou Jaoudé for having raised this problem to us and for his useful suggestions.

REFERENCES

Bosq, D. and Blanke, D. 2007. *Inference and prediction in large dimensions.* Wiley-Dunod, Chichester.

Bosq, D. and Lecoutre, J.P. 1992. *Analyse et prévision des séries chronologiques.* Masson, Paris (in French).

Brockwell, P.J., and Davis, R.A. 1991. *Time series: theory and methods, 2nd edition.* New-York: Springer-Verlag.

Hyndman, R.J. 2009. *Forecast: forecasting functions for time series.* R package version 1.24. http://www.robjhyndman.com/Rlibrary/forecast/

Mélard, G. 1990. *Méthodes de prévision à court terme.* Ellipses, Bruxelles (in French).

R Development Core Team 2008. *R: A language and environment for statistical computing* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Appendices

A1. Linear regression

First, let us examine the residuals for both periods under consideration:

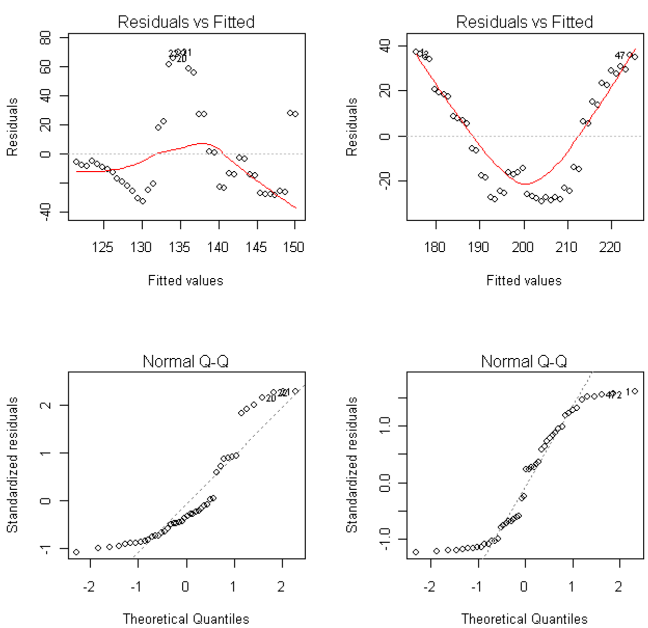

Figure A1: Residuals of the linear regression calculated from 1985-Q1 to 1995-Q4 (left) and from 1996-Q1 to 2007-Q4 (right).

The residual plots are quite unsatisfactory and present a trend. The normal QQ-plots present also irregularities in the tail values indicating that residuals have heavy tails. The obtained coefficients of determination R^2 are respectively about 0.07 and 0.28 (highlighting the lack of adequacy of the linear regression model).

Concerning the first period, we obtain that the regression coefficients are not statistically significant at level $\alpha=0.05$. For 1996-Q1 to 2007-Q4, the regression line is:

$$-8288.2947 + 4.2403\, t$$

where t represents the considered year (for example t = 1996.25 for 1996-Q2). The forecasts for 2008-2009 are:

Table A1. Forecasts by Linear Regression Calculated from 1996-Q1 to 2007-Q4

	Qtr1	Qtr2	Qtr3	Qtr4
2008	226.3054	227.3655	228.4256	229.4857
2009	230.5457	231.6058	232.6659	233.7260

For the period 1985-1995 (resp. 1996-2007), the mean-square prediction error (the mean of square differences between observed and fitted values) is large: 912.27 (resp. 547.94).

A2. Parabolic regression

Concerning the first period, the same diagnostics can be performed on the obtained residuals; the multiple coefficient of regression R^2 is still poor (of order 0.2). For 1996-Q1 to 2007-Q4, residuals are more satisfactory and, now, the R^2 is of order 0.93.

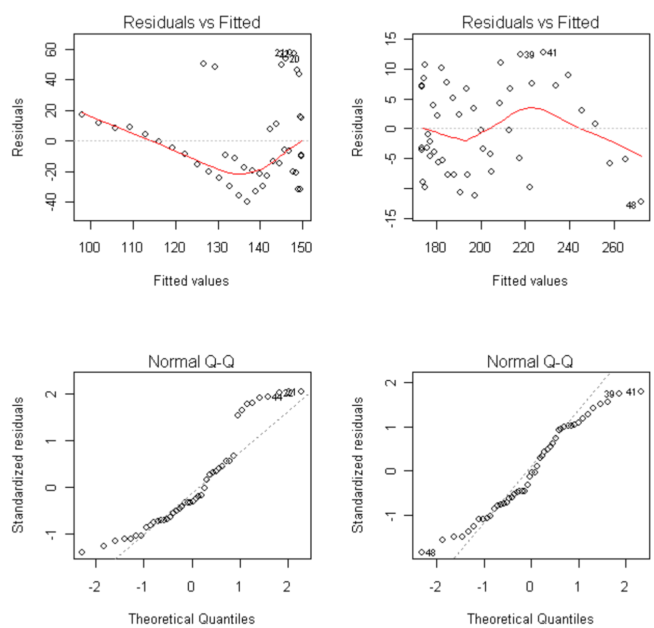

Figure A2: Residuals of the parabolic regression calculated from 1985-Q1 to 1995-Q4 (left) and from 1996-Q1 to 2007-Q4 (right).

In the last case, the obtained parabolic regression equation is:

$$8334942 - 8331.2\,t + 2.081908\,t^2$$

where t still represents the considered year. The forecasts for 2008-2009 are:

Table A2. Forecasts by parabolic regression calculated from 1996-Q1 to 2007-Q4

	Qtr1	Qtr2	Qtr3	Qtr4
2008	279.4374	287.0035	294.8298	302.9163
2009	311.2631	319.8701	328.7373	337.8648

Finally, concerning the period 1985-1995, the mean-square prediction error is equal to 787.8 while for 1996-2007, it is equal to 49.7 (which is an acceptable value compared to those obtained below for HW and BJ methods).

A3. Holt-Winters method

For 1985-Q1 to 1995-Q4, the Holt-Winters additive model is without trend but with a seasonal component. The adjusted smoothing values are $\alpha=1$, $\beta=0$ and $\gamma=1$. Starting values for the intercept a and the seasonal components s_1, s_2, s_3, s_4 are: $a=175.53125$,

$s_1=0.28125$, $s_2=-0.59375$,

$s_3=-1.65625$ and $s_4=1.96875$.

The mean-square prediction error for this model is 228.28 (better than for polynomial regression).

Concerning the period 1996-2007, the Holt-Winters adjustment performs better than the parabolic one with a mean-square prediction error of order 32.72 (instead of 49.7). In this case, the adjusted model has a trend and smoothing coefficients are $\alpha=0.7158514$, $\beta=0.3077308$ and $\gamma=1$. Starting values for a, b and s_1, s_2, s_3, s_4 now are:

$a=260.3981317$, $b=3.8470212$,

$s_1=4.4370174$, $s_2=1.0006902$,

$s_3=-3.9535466$ and $s_4=-0.3981317$.

Forecasts for 2008 and 2009 are given in the following table.

Table A3. Forecasts by Holt-Winters calculated from 1996-Q1 to 2007-Q4

	Qtr1	Qtr2	Qtr3	Qtr4
2008	268.6822	269.0929	275.8927	275.3881
2009	284.0703	284.4809	291.2808	290.7762

A4. Box-Jenkins method

For 1985 to 1995, the adjusted model by the BJ method is a SARIMA(2,0,2)$(1,0,1)^4$ with non-zero mean (using the R package "forecast", Hyndman, 2009). The computed coefficients and their empirical standard deviation are given in Table A4.

Table A4: Coefficients and their standard deviations for the period 1985-Q1/1995-Q4.

	ar1	ar2	ma1	ma2
Coefficient	0.6947	0.0314	0.3201	0.9776
SD	0.1574	0.1818	0.111	0.5624
	sar1	**sma1**	**intercept**	
Coefficient	-0.3237	0.7673	141.9635	
SD	0.3396	0.214	16.0001	

As expected, the mean square prediction error is about 104.68, better than for the previously tested models. The quality of the fit is measured through the following diagnostics which show that the error term of the model can be considered as white noise (see Figure A3).

Concerning the period 1996-Q1/2007-Q4, one obtains a ARIMA(2,2,1) with estimated coefficients given in Table A5. The mean-square prediction error is now about 23.97 (instead of 32.72 for the Holt-Winters method).

Table A5. Coefficients and their standard deviations for the period 1996-Q1/2007-Q4

	ar1	ar2	ma1
Coefficient	-0.1371	0.4539	-0.8841
SD	0.1847	0.1777	0.1057

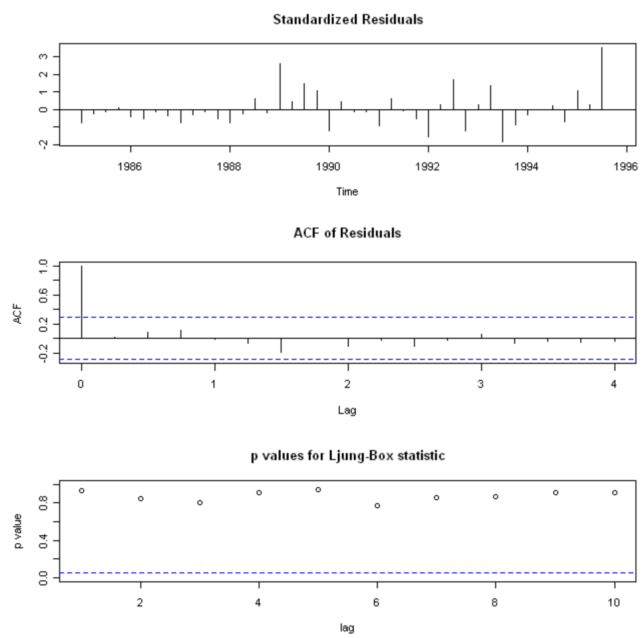

Figure A3: BJ residuals. Period 1985-Q1/1995-Q4

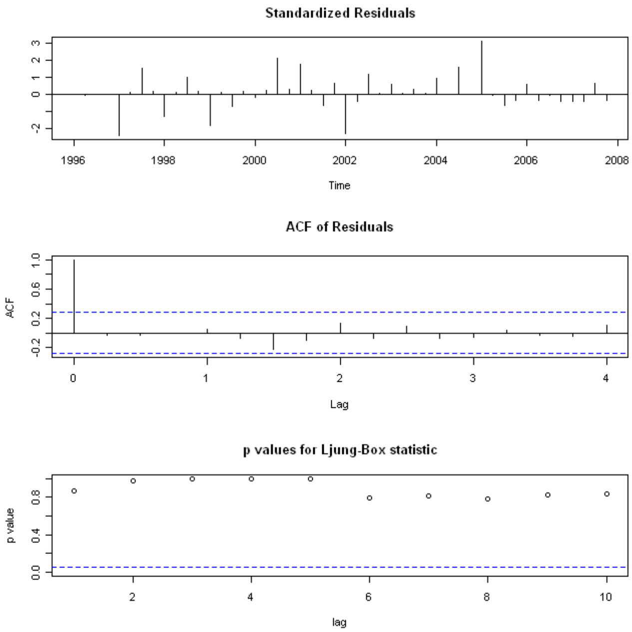

Figure A4: BJ residuals. Period 1996-Q1/2007-Q4

Again, the error term of the model can be considered as white noise (see Figure A4). Finally we give in Table A6, the prediction interval at confidence level 80% (resp. 95%):

[Lo 80 , Hi 80] (resp. [Lo 95 , Hi 95])

while "Forecast" corresponds to the fitted value.

Table A6. Forecasts and their prediction intervals for 2008-2009

Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2008:					
Q1	265.8799	259.4706	272.2891	256.0778	275.6820
Q2	267.5490	258.5805	276.5175	253.8328	281.2651
Q3	272.4645	259.0733	285.8558	251.9844	292.9447
Q4	275.0236	258.2369	291.8102	249.3506	300.6965
2009:					
Q1	279.3794	258.4455	300.3133	247.3637	311.3950
Q2	282.4191	257.7748	307.0634	244.7289	320.1093
Q3	286.4549	257.7276	315.1821	242.5203	330.3894
Q4	289.7567	257.1195	322.3938	239.8425	339.6708

A5. Nonparametric kernel prediction

Coming back to CPC prices, we begin by interpreting the BJ results of Annex A4. Since $I = 2$ for 1996-Q1/2007-Q4, we construct our nonparametric predictor on the twice differentiated data (Y_t) which can be considered as stationary:

$$Y_t = X_t - 2X_{t-1} + X_{t-2}$$

where t is varying from 1996-Q3 to 1997-Q4 (so that n=46). Moreover, using Table A4 and inverting the process, we find that the best linear predictor of Y_{n+1} given $Y_t, t \le n$ is

$$0.74Y_n - 0.2Y_{n-1} + 0.17Y_{n-2} - 0.15Y_{n-3} + \cdots$$

Noting the important role played by Y_n , it is convenient to use the couples (Y_i, Y_{i+H}) for computing the NW predictor. This allows us to avoid the well-known nonparametric curse of dimensionality. Values obtained for $\hat{Y}_{2008-Q1}, \hat{Y}_{2008-Q2}, \hat{Y}_{2008-Q3}$ and $\hat{Y}_{2008-Q4}$ are given in Table A7. Predictors of $\hat{X}_{2008-Q1}, \hat{X}_{2008-Q2}, \hat{X}_{2008-Q3}$ and $\hat{X}_{2008-Q4}$ can be computed from the equations:

$$\hat{X}_{2008-Q1} = \hat{Y}_{2008-Q1} - 2X_{2007-Q4} + X_{2007-Q3}$$
$$\hat{X}_{2008-Q2} = \hat{Y}_{2008-Q2} - 2\hat{X}_{2008-Q1} + X_{2007-Q4}$$
$$\hat{X}_{2008-Q3} = \hat{Y}_{2008-Q3} - 2\hat{X}_{2008-Q2} + \hat{X}_{2008-Q1}$$
$$\hat{X}_{2008-Q4} = \hat{Y}_{2008-Q4} - 2\hat{X}_{2008-Q3} + \hat{X}_{2008-Q2}.$$

We compute the Nadaraya-Watson estimator by choosing a standard normal kernel. Finally, the bandwidth h_n is empirically adjusted by minimizing the prediction error for the observed year 2007. This method leads to forecasts that are slightly lower than values obtained by the BJ method (see Table A5) but still within the prediction intervals.

Table A7. Forecasts for 2008 by NW predictor

	2008-Q1	2008-Q2	2008-Q3	2008-Q4
$\hat{Y}$	4.424399	-3.337333	3.203995	-1.960531
$\hat{X}$	264.3613	265.3628	269.6254	271.6383

Correspondence: delphine.blanke@univ-avignon.fr