# Modélisation et prévision des prix du CPC

Delphine Blanke, Denis Bosq

# MODELISATION ET PREVISION DES PRIX DU CPC

**D. Blanke**
*Université d'Avignon et des Pays de Vaucluse, France*
**D. Bosq**
*Université Pierre et Marie Curie - Paris 6, France*

***Résumé :*** *Dans cet exposé, nous étudions la modélisation des prix du CPC (coke petroleum calcination) et utilisons les modèles obtenus pour la prévision. La difficulté du problème vient de la non-homogénéité des observations.*
*Nous considérons d'abord des techniques empiriques (régression polynomiale) et nous remarquons qu'elles sont des cas particuliers de la méthode SARIMA. Nous utilisons alors la méthode SARIMA et nous la comparons à une méthode non paramétrique.*

***Mots-Clés :*** *Modèle ARMA,Prévision, Non paramétrique, Régression polynomiale.*

***Abstract :*** *In this talk we model and predict the time series of coke petroleum calcination prices. Intricacy of this problem comes from the nonhomogeneity of data.*
*We first consider empirical technics like polynomial regression and notice that they appear as special cases of the SARIMA method. Then, we use the SARIMA method and compare it to a nonparametric method.*

***Keywords :*** *ARMA model, Forecasting, Nonparametric, Polynomial regression.*

## 1. Introduction

The origin of the current study is a request concerning a project of privatization of a Company, requiring its evaluation. Current market conditions for CPC are involved in the future turnover of this project. The CPC (Coke Petroleum Calcination) is derived from a by-product of oil and it is used in aluminum and titanium alloys. In this paper, we study structure and prediction of CPC price (in dollars) from quarterly data (denoted from Q1 to Q4) between 1985-Q1 and 2008-Q4. 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), Box and Jenkins method (BJ) and finally, nonparametric prediction (NP). For this purpose, we have used the software *R* actually developed by the R Development Core Team (2008).

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.

Concerning the data, their non-homogeneity led us to cut them in 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 for evaluation of the company).

The next section deals with polynomial regression and HW. We specify their link 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:

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

where $(X_t)$ is the observed process, $a_0$ and $a_1$ are real coefficients and $(\varepsilon_t)$ is a white noise:

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

We first show that, in some sense, the LR model is a special ARIMA process (ARIMA theory appears in Brockwell and Davis (1991) among others). Set

$$Y_t = X_t - X_{t-1} - a_1, \quad t \in \mathrm{Z} \qquad (1)$$

then

$$Y_t = \varepsilon_t - \varepsilon_{t-1}, \quad t \in \mathrm{Z} \qquad (2)$$

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, $(\varepsilon_t)$ is the innovation process of $(Y_t)$. In order to prove that assertion, we first note that (2) implies

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

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 $\varepsilon_t$ is orthogonal to $M_{t-1}$, we conclude that $(\varepsilon_t)$ 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 correlation between the $X_t$'s. This lack is clear since the "explained variances" $R^2$ are respectively 7% and 28% !

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

$$X_t = a_0 + a_1 t + a_2 t^2 + \varepsilon_t, \quad t \in \mathrm{Z}. \qquad (4)$$

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

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

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 $(\varepsilon_t)$ is the innovation of $(Y_t)$. In order to prove that claim, we set

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

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.

Now the PR suffers for the same lack as LR but it is more adapted to CCP from 1996 to 2007 with a $R^2$ of 93%. 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 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 $X_1,...X_T$, 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 $\alpha$, $\beta$ and $\gamma$ are taken in [0,1]. Their optimal values are determined by minimizing the squared one-step prediction error. 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 (see, for example, Bosq and Lecoutre, 1992). The obtained results by this method are quite satisfactory for both periods.

### 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 is one year and there is no trend. However, the forecasts for 1997 and 1998 are not completely satisfactory, highlighting the change of model. Concerning 1996-2007 the new model is ARIMA (2,2,1). 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 the talk. 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, remark that the model remains the same by considering only the data 1996-Q1 to 2006-Q4 leading to sharp forecasts of the observed year 2007. On the contrary, if one adds the two first data of 2008, the model becomes a ARIMA (1,2,0) but with a less good adjustment.

### 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 $1,...,n$, 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 $i = 1,...,n-H$. 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 $f(x,y)$ not depending on $i$, then

$$E(Y_{i+H} / Y_i = x) = \int_{-\infty}^{\infty} y f(x,y) \mathrm{d}y \Big/ \int_{-\infty}^{\infty} f(x,y) \mathrm{d}y. \qquad (4)$$

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

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

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

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

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

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

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

Basing 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.

## 5. Conclusion

We have seen that the LR, PR and HW methods are all associated with the 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.

## Bibliography


Blanke, D. and Bosq, D. *Modeling and Forecasting the CPC Prices.* To appear in CSBIGS, 2010.

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

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

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

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

Mélard, G. *Méthodes de prévision à court terme.* Ellipses, Bruxelles, 1990.

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