Estimation of Volatility of Cross Sectional Data: a Kalman filter approach

Cristina Sommacampagna[*]
University of Verona
Italy

Gordon Sick[†]
University of Calgary
Canada

This version: 4 April, 2004

Abstract

In order to perform a Real Option evaluation some variables have to be estimated. One of the main variables to estimate is the volatility of the variable underlying the option. With respect to the financial options, because of the different objects of the evaluation, the available data can need a specific kind of analysis.

In this document we consider the case of a crude oil producing company that collected data about the past costs of drilling wells and now wants to evaluate the real option of starting drilling new wells: we need to estimate the volatility of the drilling costs.

Keywords: Real Options; Cross-Sectional Data; Kalman Filter.

1 Brief Introduction

A real option differs from a financial option because of the underlying variable and because it can provide an evaluation of the upside potential of a real project; so leading to an optimal strategy in terms of timing of development, Sick [5] and [6]. In particular, when we consider the option of developing a new project that requires an investment (both initially or during the life of the project), an important element of the analysis is given by the volatility of the cost of development.

[*] PhD Student in "Mathematics for Economic Decisions", Department of Economics. On the web: *http://web.economia.univr.it/safe* selecting *Staff*. Email: *cristina_sommacampagna@pilar.univr.it*.

[†] Professor of Finance, Haskayne School of Business. On the web: *http://www.ucalgary.ca/∼sick*. Email: *sick@ucalgary.ca*.

The main source of data for the cost of development of a project is probably given by the companies involved in that kind of industry. In this case, especially if we are dealing with one company only, the data set has peculiar characteristics and a peculiar analysis is required.

We consider the case, quite frequent, in which the available data do not show the same frequency within each month of observation and are not always generated by the observation of the same phenomenon.

The next section illustrates in detail the characteristics of the problem we are facing; section 2 focuses on the characteristics of the analyzed data; section 3 illustrates the characteristics of the technique we decide to use, given the problem we want to solve; finally section 4 shows the results of our estimations.

2 Data Description

The object of our study is the data-set collected by a Petroleum Company[1] from January 1990 to March 2000. The database we use reports, for every well drilled by the company: the serial identification number, the drilling starting and ending or abandoning date, the total cost of drilling, the geographic location, the type and the total length of the perforation.

From the available data we can calculate, for each well, both a cost per foot and a cost per day of drilling. Since the data are referring to different kind of wells - we can in fact distinguish between vertical and horizontal wells, directional and re-entry wells - we decide to refer to the cost per day of drilling, since the comparison of costs per food can be misleading when different kind of wells are considered. We refer to the cost per day of drilling as the *drilling day rate*: we calculate it as the ratio between the total cost of drilling a specific well and the number of days necessary to complete the perforation.

The extraction activity, as results from the database, concern many different areas. Nevertheless we concentrate our first analysis on the data referred to one of this areas since it generates 1413 observations for 123 months, more than 50% of the available data.

The data that we analyze have three main characteristics:

- first of all, every observation in the data series concerns a different well; therefore the costs are not homogenous and part of the volatility of the registered costs depends on the different characteristics of every single well;

- secondly, what we observe are the costs of drilling wells that have been finally chosen by the managers of the company; we have no knowledge of the true process that drives the costs;

[1] The source of the data cannot be disclosed for privacy reasons.

- finally, the observations are distributed on an horizon of more than nine years, the average time for drilling a well is one month, but we have more than one observation for every month.

From the first characteristic arises that we are facing some sort of panel data and one of our objective must be that of estimating a volatility that that does not include the volatility due to the different characteristics of every considered well. From the second characteristic comes the necessity of using some particular technique for the variance estimation of an hidden process. Finally, the last characteristic suggests that we can use the available data on a monthly base only if we can take into account that every month we have a different number of observations.

Figure 1 shows the number of wells drilled in the considered region in every month of the considered period.

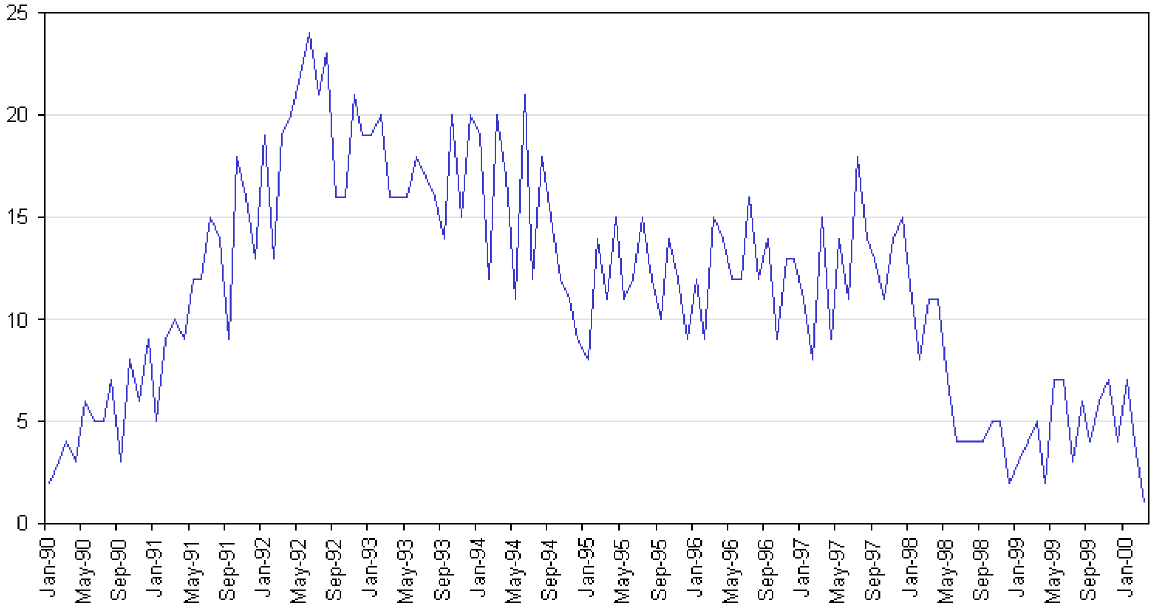

Figure 1: Number of drilled wells per month in the considered area.

Figure 2 reports the scatter plot of the drilling day rates on the months of observation; the line is the average per month of the same drilling day rates. There are some evidences:

- first of all, from August 1996 the dispersion around the mean is much higher and the cost level shows a pick in the period August 1996 - December 1998; after that however, the average seems to go back to the preceding level;

- secondly, the costs clearly show to be cyclical in the period that approximately goes from May 1991 to December 1997; this coincide with the period in which more observations are available;

- finally, some outliers are present for the dates of May 1990, October 1991 and January 1993, we do not consider this values for the estimation.

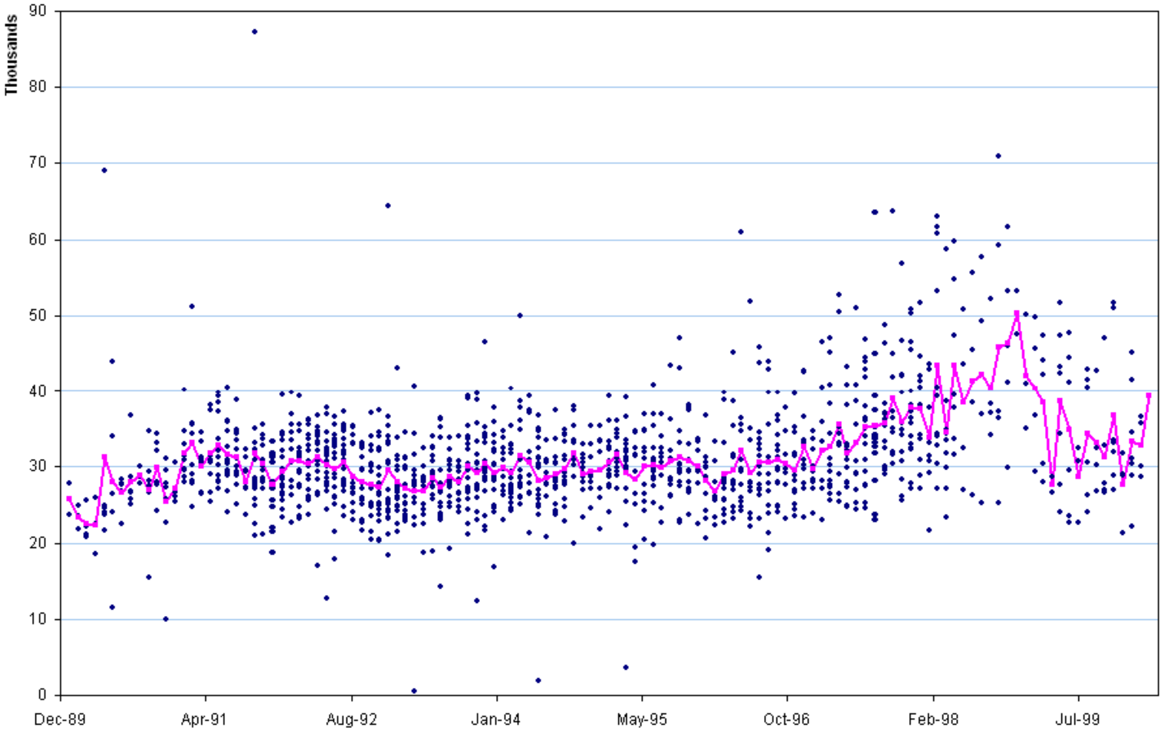

Figure 2: Cost of drilling per day for every well in the considered region: 1413 data points from Jan-1990 to Mar-2000. The line is the average drilling day rate per month.

Finally, Figure 3 reports the length in feet of the vertical section of each well drilled in the considered region, together with the rescaled[2] cost per foot of perforation. As point out by the circle in the graph, during the period Jan-1998 to Jul-1999 the wells are characterized by a much smaller depth and at the same time by a quite high volatility in cost per feet of perforation. Refereing back to Figure 2 we notice that in the same period also the drilling day rate is characterized by a higher volatility: we can deduce that, during this period, the firm is drilling wells that implies bigger difficulties, probably because of the characteristics of the ground.

In Figure 4 we plot the length of the vertical sections alone to have a more readable graph. We can notice that the history of drilling for the firm is definitely increasing, except for the already cited period Jan-1998 to Jul-1999. We deduce that the firm had the possibility to chose, among wells, to drill where the perforation was presenting less difficulties, leaving last the wells that required a higher cost for a small length of perforation.

[2] We rescaled the cost per foot of perforation multiplying the vector by a factor of 10, to have a readable graph

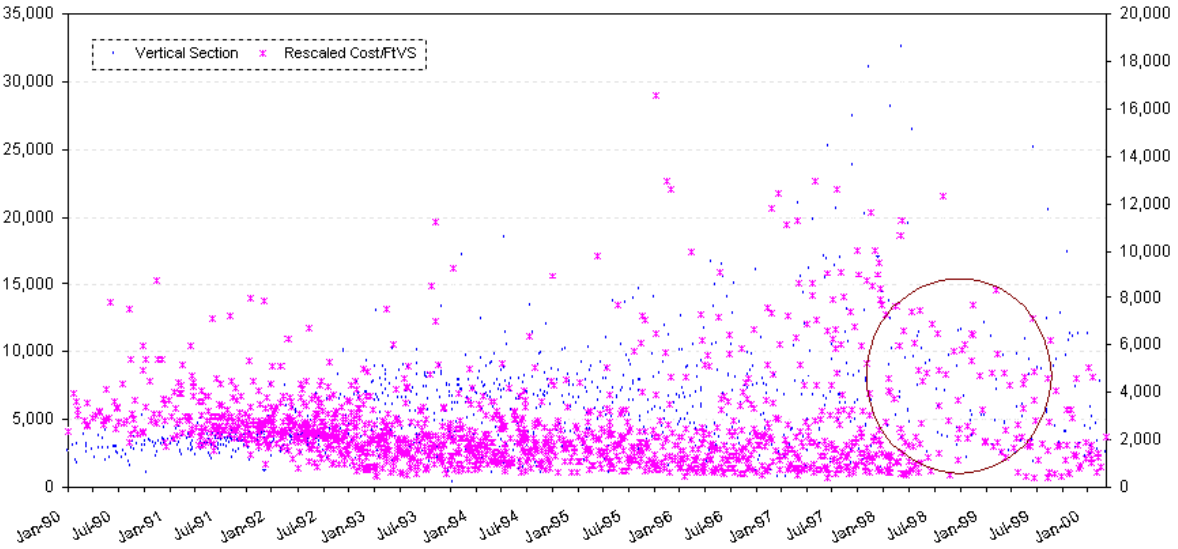

Figure 3: Vertical section depth and cost per foot (rescaled multiplying by a factor of 10) of drilling, for every well in the considered region: 1413 data points from Jan-1990 to Mar-2000.

3 Estimation of Volatility with the Kalman Filter

To estimate the volatility of the drilling day rate we need to solve the particular estimation problem that arises from the characteristics of the data we are considering; we develop the following approach: we model the process for the cost of every single well as function of a true underlying cost process. We give to the underlying cost process, which can be defined as *hidden*, the form of an auto-regressive process with trend μ.

The state-space representation of the dynamics of this model, in a general formulation, is given by the following system of equations:

$$C_{i,t} = Z_t c_t + d_t + \varepsilon_{i,t} \tag{1}$$

$$c_t = \mu c_{t-1} + a_t + \xi_t. \tag{2}$$

Where $C_{i,t}$ is the cost registered for every single well i in the month t; c_t is the hidden cost at time t. d_t and a_t are the intercepts of the model; $\varepsilon_{i,t}$ and ξ_t are vectors of white noises realizations with mean zero and matrix of covariance respectively H_t and Q_t. For the generic element of the row r we have

$$E(\tilde{\varepsilon}^r_{i,t}\tilde{\varepsilon}^r_{j,t}) = \begin{cases} 0 & \text{if } i \neq j \\ \sigma^2_{h,t} & \text{if } i = j, \end{cases} \tag{3}$$

$$E(\tilde{\xi}^r_t\tilde{\xi}^r_\tau) = \begin{cases} 0 & \text{if } t \neq \tau \\ \sigma^2_{q,t} & \text{if } t = \tau. \end{cases} \tag{4}$$

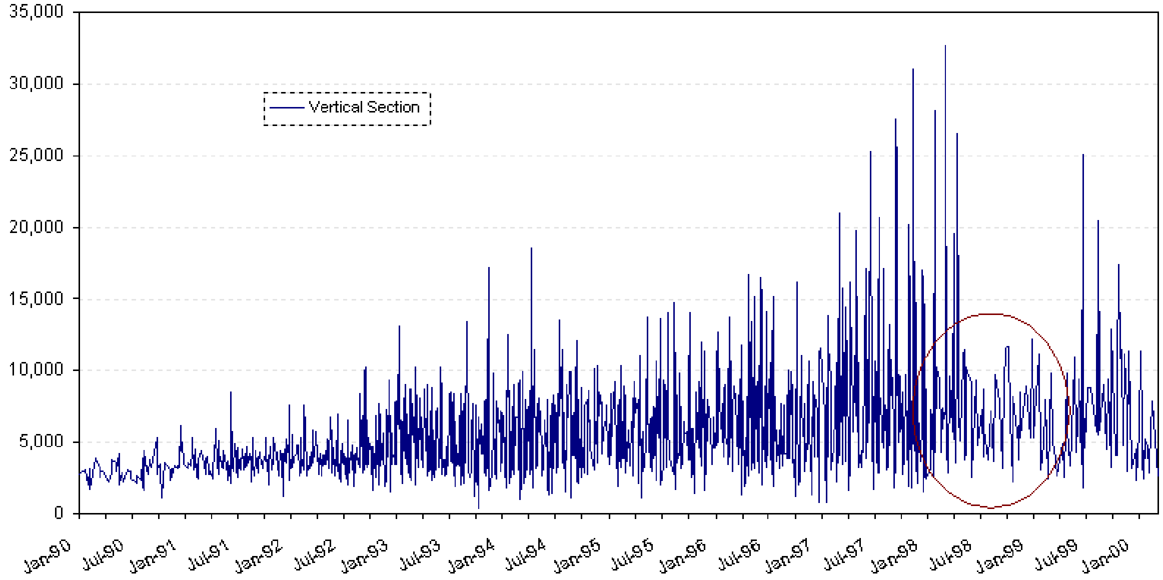

Figure 4: Vertical section depth of drilling, for every well in the considered region: 1413 data points from Jan-1990 to Mar-2000.

To carry the estimation of the volatility, crucial for our analysis, we observe that the equations 1 and 2 form a state space representation of our model, where the first equation is known as *observation equation* and the second equation is known as *state equation*; therefore, to estimate the model, we use the Kalman filter technique. The Kalman filter, is a recursive algorithm which allows one to upgrade the model estimates using new information and to estimate the value of an unobservable variable, Hamilton [3] and Harvey [4].

Through the Maximum Likelihood technique, the Kalman filter returns the estimation of the underlying cost value c_t, for every time t, of all the parameters Z_t, d_t, μ and a_t, of the covariance matrixes of the two error terms.

For our application we mainly need to determine the covariance matrix Q.

3.1 Standard Kalman Filter

What follows is a general description of the Kalman filter, as described in from Harvey [4] and Hamilton [3].

The state space representation is given by observation (or measurement) equation and by state (or transition) equation, whose dimension is specified in the following expressions:

$$y_t^{(N\times 1)} = Z_t^{(N\times m)} x_t^{(m\times 1)} + d_t^{(N\times 1)} + \varepsilon_t^{(N\times 1)}, \tag{5}$$

$$x_t^{(m\times 1)} = T_t^{(m\times m)} x_{t-1}^{(m\times 1)} + a_t^{(m\times 1)} + \xi_t^{(m\times 1)}, \tag{6}$$

where y_t is the vector of observations and x_t is the vector of unobservable state variables; Z_t is a matrix of known or unknown, constant or time varying coefficients; matrix T_t is the state transition matrix; d_t and a_t are known or unknown vectors. Finally ε_t is a vector of serially uncorrelated disturbances with mean zero and covariance matrix H_t, and ξ_t is identified with a vector of serially uncorrelated disturbances with mean zero and covariance matrix Q_t. The unobservable process, the elements of the matrices and the variances of the noise processes, are estimated by maximizing the likelihood function

$$\log L = -\frac{NM}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{M}\log F_t - \frac{1}{2}\sum_{t=1}^{M}\frac{v_t' v_t}{F_t} \tag{7}$$

where v_t is the one-step ahead residual at time t, $v_t = y_t - \hat{y}_t$, F_t is its variance, N is the length of the vector of observations y_t and M is the length of the vector of the unobservable state variables. The system is recursively updated using the following equations

$$x_{t|t-1} = Tx_t + a \tag{8}$$

$$P_{t|t-1} = TP_tT' + Q \tag{9}$$

$$v_t = y_t - Z_t x_{t|t-1} - d \tag{10}$$

$$F_t = Z_t P_{t|t-1} Z_t' + H \tag{11}$$

$$x_t = x_{t|t-1} + \frac{P_{t|t-1} Z_t' v_t}{F_t} \tag{12}$$

$$P_t = P_{t|t-1} - \frac{P_{t|t-1} Z_t' Z_t P_{t|t-1}}{F_t}. \tag{13}$$

Equations 8-13 are known as the Kalman filter. They are a series of regressions that generate an estimate of the state vector, x_t, and its covariance matrix, P_t. Given estimates of the starting values of these two variables, x_0 and P_0, an estimate of the unknown regression coefficient and then parameters of the model are derived.

Once these estimates have been obtained, we have an estimate of the state vector, the recursive residuals and their variance, and can also generate an estimate of the updated residual vector $e_t = y_t - Z_t x_t - d_t$.

3.2 Kalman Filter for Incomplete Panel Data

In our particular setting, we apply the Kalman filter for incomplete panel-data, as in Harvey [4]. To apply the Kalman filter in fact it is not necessary to have the same number of observations at each time t.

The observation and state equations do not change with respect to the standard formulation given in equations 5 and 6 but the dimension of the vectors y_t, x_t, d_t, ε_t and of the matrix Z_t change with time:

$$y_t^{(N_t \times 1)} = Z_t^{(N_t \times m)} x_t^{(m \times)} + d_t^{(N_t \times 1)} + \varepsilon_t^{(N_t \times 1)}, \tag{14}$$

$$x_t^{(m \times 1)} = T_t^{(m \times m)} x_{t-1}^{(m \times 1)} + a_t^{(m \times 1)} + \xi_t^{(m \times 1)}. \tag{15}$$

The covariance matrix of the error term ε_t also change with time (H_t).

In a recent paper Cortazar, Schwartz and Naranjo [2] apply this technique to the estimation of current term structure of interest rates and its dynamic, having sparse bond prices.

We use this technique since we have a different number of observations for every month; we interpret them as different realizations of the same underlying process.

4 Results

In our application, y_t is the drilling day rate and x_t the hidden cost. We fix Z_t to be a vector of ones, d_t to be a vector of zeros, T_t to be zero and a_t to be zero; the matrix of covariance of the error terms are respectively: the matrix $H_t^{(N_t \times N_t)}$ and $Q_t^{(1 \times 1)}$, which is the parameter objective of this application. An important assumption is that the matrix H is diagonal and all the elements on the diagonal are equal.

The unobservable process c_t, the variance of this estimation, the coefficients of the model, the covariance matrixes of the noise processes, are estimated maximizing the likelihood function shown in equation 7.

The estimation is run using a moving window of 24 months (even if this means to consider the two preceding years of observation, it allows to avoid some of the numerical problems we had to face); this window in fact is moved one step ahead maintaining the same dimension. The procedure results in 99 estimations of the unobservable cost and of variances of the error terms.

In figure 5 we report the estimation of the unobservable costs (blue line), obtained with the Kalman filter estimation technique, for incomplete panel data. The same figure also reports the average of costs per month (pink line) as a term for comparison. The results go from January 1992 to March 2000.

The estimated costs show a much smoother dynamic with respect to the average per month of costs. This is due to the fact that the Kalman filter identify which part of

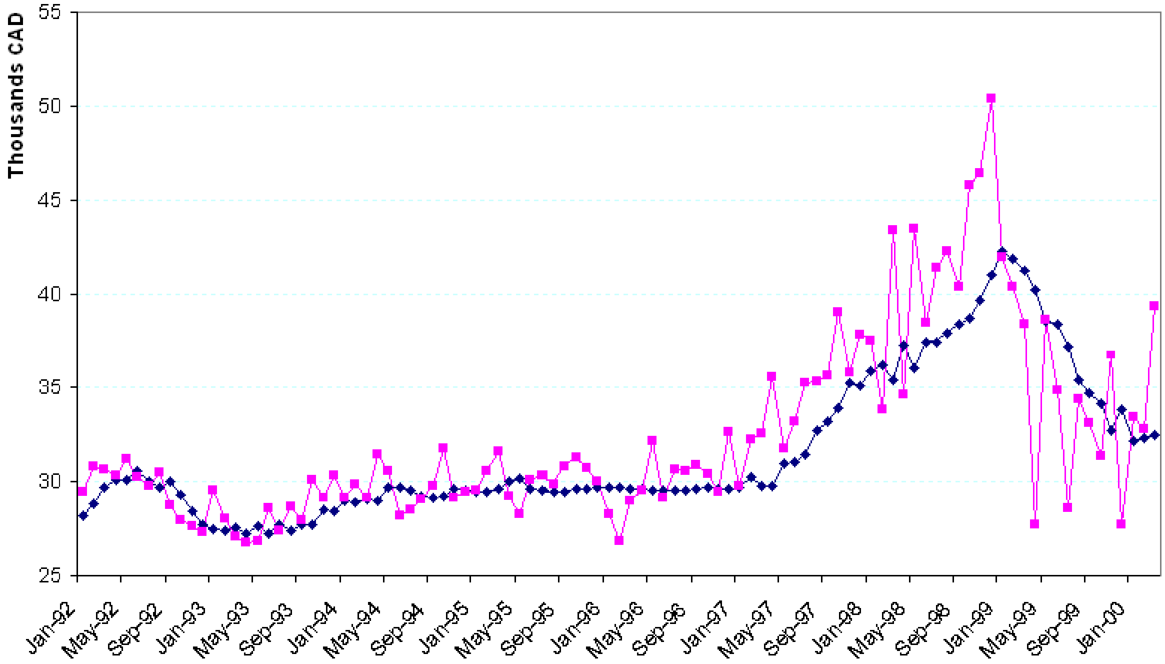

Figure 5: The blue line shows the unobservable cost estimated with the Kalman filter for incomplete panel data; the pink line reports the monthly average of costs.

the variation in the data is to be attributed to the observation process and which part is to be attributed to the underlying process. More, since we adopted the procedure for incomplete panel data, we can interpret the consistent reduction of variance to the attribution of part of the volatility to the variation among the data grouped per month, which represent different realizations of the phenomena.

In figure 6 the pink line represent the estimation of the volatility σ_q of the error term of the state equation (as defined in equation 4); the blue line is the volatility of the estimated parameter for the hidden cost. This volatilities and the one presented in the following figure are monthly volatilities.

It is clear from the graph that there are problems of estimation for the sub period May 1995 - May 1997. However we can notice that, out of this sub-period, the volatility of the hidden process is much lower than the volatility of the average per month of the drilling costs that moves around the value of 16.69% on a monthly base.

Finally, figure 7 shows the variance σ_h of the error term for the measurement equation. This is considerably higher than the volatility of the state equation, making think to the need of better specifying the model.

The estimation has been carried in Matlab environment.

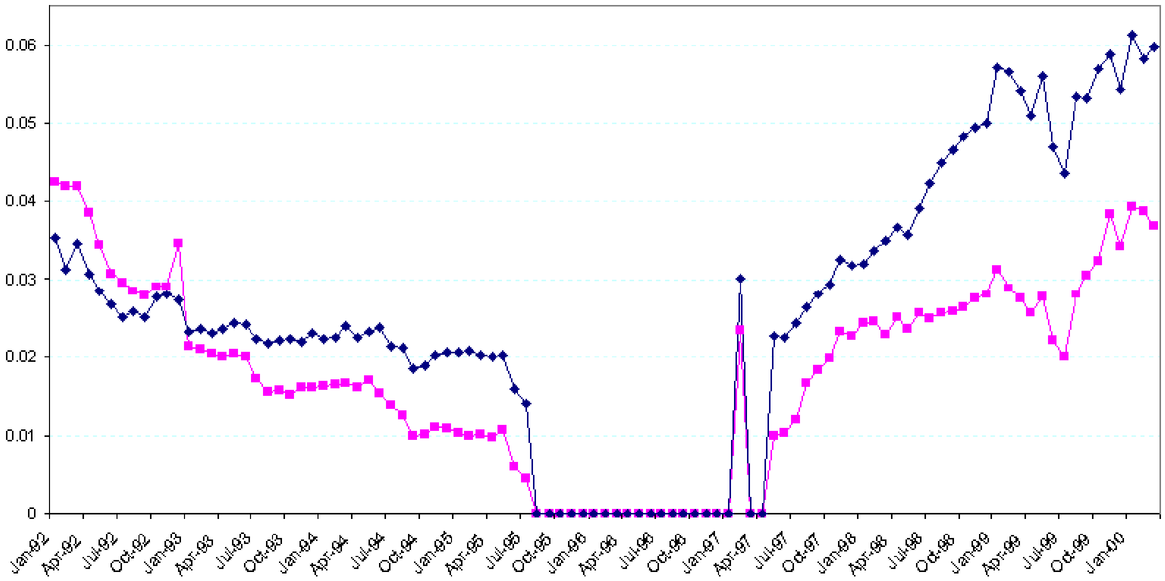

Figure 6: The pink line shows the estimated volatility of the error term of the state equation; the pink line reports the volatility of the estimation of the hidden cost.

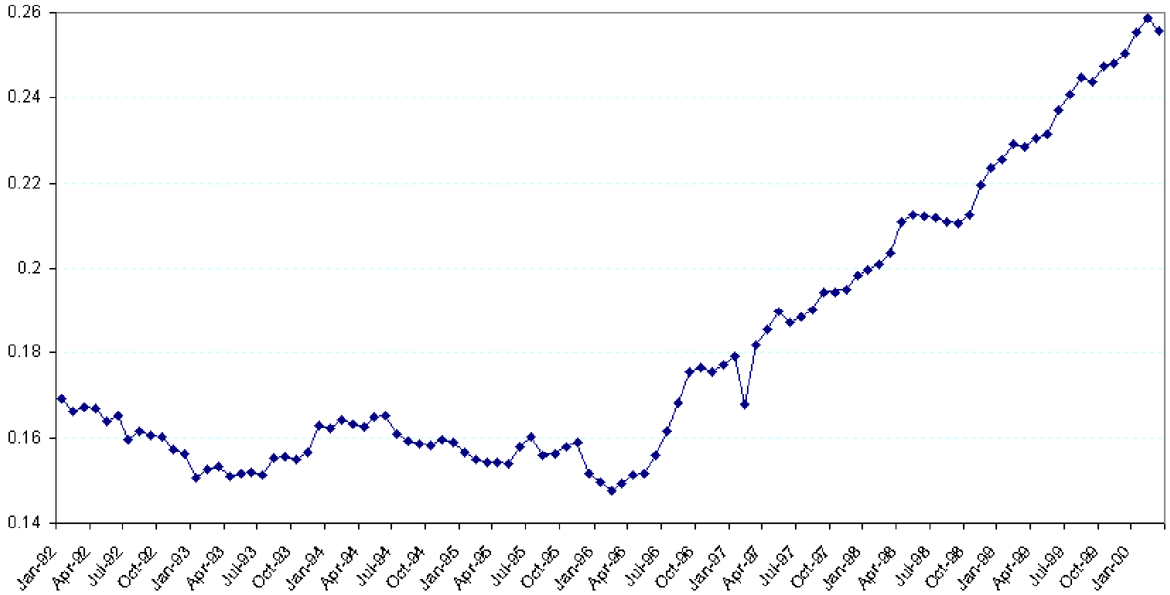

Figure 7: estimated volatility of the error term of the measurement equation.

5 Further Developments

The first targets are: generate an estimation with no numerical problems, maybe developing this approach in a different programming environment; test different specifications for the model of costs.

Second, from figure 2 we notice that there probably is a seasonal component in the data we are using. The next target will be that of modifying the state space representation in order to take seasonality into consideration.

Finally, we would like to develop a model of investment for the considered company, in order to verify the goodness of the fit of the estimated data as instruments to determine the policy of the company.

References

[1] Baltagi, B.H., *Econometric Analysis of Panel Data*, John Wiley&Sons Ltd, New York, 2001.

[2] Cortazar, G., Schwartz, E. and Naranjo, L. "Term Structure Estimation in Low-Frequency Transaction Markets: A Kalman Filter Approach with Incomplete Panel-Data", March 2003.

[3] Hamilton, J. (1994), *Time Series Analysis*, Princeton University Press, Princeton, NJ.

[4] Harvey, A. C. (1994), *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, Cambridge.

[5] Sick, G. (1989), "Capital Budgeting with Real Options", printed in *Monograph 1989-3*, Series in Finance and Economics, Salomon Brothers, Center for the Study of Financial Institutions, Leonard N. Stern School of Business, New York University.

[6] Sick, G. (1995), "Real options", published as Chapter 21 of *Finance*, Ed. R.A. Jarrow, V. Maksimovic, W.T. Ziemba, *Handbooks in Operations Research and Management Science*, Volum 9, Elsevier North Holland.