Combining quantitative and qualitative data in business cycle research

E. Smit
Department of Statistics, University of the Western Cape, Bellville

In this article a leading indicator of the South African business cycle is proposed which combines the traditional quantitative data inputs with qualitative data. The integration is achieved via the Kalman filter technique. It is shown that this model surpasses the traditional approaches in accuracy.


Introduction

It may be too much of a generalization to state that a schism exists between the protagonists and users of quantitative and qualitative data respectively. However, the main thrust of business cycle research in the United States has been directed towards the construction of quantitative indicators, whereas the accent in Europe has been on the analysis of qualitative survey data. Efforts directed towards a synthesis were mainly due to econometricians trying to improve the forecasting accuracy of models based on quantitative data by introducing qualitative variables into their equations. Our own efforts have shown that quantitative variables tend to suppress qualitative variables when used in combination and that qualitative variables tend to suppress each other due to high levels of multicollinearity (Smit, 1983). On balance, qualitative data have the advantages of early availability, of capturing expectations data, and of providing information on a broader spectrum of variables than traditionally published in quantitative form, whereas quantitative data excel providing 'richer' or 'fuller' information due to measurement on a higher measurement scale. The problem is how to optimally integrate and use both types of data in economic analyses and forecasting. It is believed that the Kalman filter technique opens new horizons in this regard.

In 1975 Aiginger used the Kalman filter to illustrate how quantitative investment anticipations could be used to correct constants in the investment equation (Aiginger, 1975). This paper reports on research aimed at incorporating qualitative data into a quantitative model via the Kalman filter. The problem is the establishment of a leading indicator for the South African business cycle. Two approaches are described: a quantitative technique used by Barr (1983) and a qualitative technique after Smit & Kotze (1984). The Kalman filter is then used to integrate these approaches.

Measuring the business cycle: van Collar's coinciding indicator

Since 1977 the Bureau for Economic Research at the University of Stellenbosch has been publishing composite leading and coinciding indicators of the South African business cycle on a monthly basis. These indices were developed by van Coller (1980) using the methods of the NBER. The five series constituting the coinciding indicator are as follows: registered unemployment, index of physical volume of manufacturing production, wholesale sales, retail sales, and imports.

This composite coinciding indicator has proved to be of
great value in the South African context and many firms use this variable as an input in their own forecasting models, justifying the effort exerted in forecasting the coinciding indicator. The leading indicator developed by van Coller, speaking on the average, has sufficient lead time to be a truly ex-ante indicator, but it has the disadvantage of inconsistent lead times at peaks and troughs ranging between 3 and 21 months. This has led to efforts to establish an alternative leading indicator showing more consistent timing at peaks and troughs.

**An alternative based on quantitative data: Barr’s proposal**

Barr (1983) applied a stepwise regression method for selecting, in an optimal way, from 23 variables considered, those variables which represent the best set of predictors for van Coller’s coinciding indicator. This method, he argued, obviates the need to make ex-ante predictions about the relevance of the variables as leading indicators or to analyse consistency of turning-point relations because the variable with the highest correlation will necessarily provide the best average correlation with the business cycle and thus be the most consistent measure of turning points. Furthermore, this method provides weights which are automatically selected in an optimal way and hence obviates the selection of subjective weighings when averaging. The model is fitted over the period 1967 M1 – 1981 M10 and the following explanatory variables contribute towards a $R^2$ value of 0.95: new motor cars licenced, imports, commercial bank loans and discounts, and commercial bank demand deposits. Due to a lag of three months between the dependent and all of the independent variables it is possible to calculate a predicted value for the coinciding indicator three months ahead of its realization. However, due to delays in the publication of quantitative data, data on the variables at time $t$ may only become available after another four months have elapsed, destroying the true ex-ante character of the leading indicator.

**An alternative based on qualitative data: Smit and Kotze’s proposal**

Smit and Kotze (1984) stressed the importance of the psychological disposition of businessmen and argued that the expectations and plans of economic agents provide the link between future and present economic activity. They showed that psychological links do not only occur in the so-called psychological business cycle theories, but that they also feature within the so-called real theories, although not in the same exalted position. The underlying idea is that if changes in moods and expectations should lead to changes in real activity, then in theory it should be possible to predict the development of the business cycle when measurements on psychological variables are available.

They tested this hypothesis using the subjective qualitative expectations and plan data published by the Bureau for Economic Research at the University of Stellenbosch as part of the Opinion Survey and Building Survey results. All ex-ante variables related to expectations (statements about future events outside the direct sphere of influence of the firm) and plans (statements about the firm’s own future course of action) are included in the analysis provided that sufficient observations are available. These qualitative series, when compared to the quantitative series usually used in indicator research, have the disadvantages of being available only on a quarterly instead of a monthly basis and of covering only a few sectors in the economy. Furthermore about 50% of the series only date back to 1974:4. However, these series have the major advantage of being available three months ahead of the event they relate to. The following variables have qualified for selection:

- **MSales**: manufacturing; expected volume of sales;
- **Mprod**: manufacturing; expected volume of production;
- **Munor**: manufacturing; expected ratio of unfilled orders to demand;
- **Mrawmat**: manufacturing; stocks of raw materials relative to planned production;
- **Mfinpro**: manufacturing; stocks of finished goods relative to expected demand;
- **MBuscli**: manufacturing; expected business climate;
- **Wsales**: wholesale; expected volume of sales;
- **Wstock**: wholesale; stocks relative to expected demand;
- **Wbuscli**: wholesale; expected business climate;
- **Rsales**: retail; expected volume of sales;
- **Rstock**: retail; stocks relative to expected demand;
- **Rbuscli**: retail; expected business climate;
- **Cstock (N)**: motor trade; stocks of new cars/used cars/spares relative to expected demand;
- **Cbuscli**: motor trade; expected business climate;
- **Bpcap**: manufacturers of building materials; expected production capacity;
- **Bbuscli**: builders; expected business climate.

The variables are measured relative to the same quarter in the previous year (higher, the same, lower or better, the same, worse) with the exception of the variables related to stocks which are reported as too high, just sufficient or too low. The qualitative measurements are transformed to balance series on the interval [0, 100]. It has been argued that it is unnecessary to make ex-ante evaluations about the economic significance of the variables as leading indicators or to analyse the consistency of turning-point relations when a stepwise regression approach is used. Nor is it necessary to select subjective weighings when averaging because these will also be selected optimally. Similar arguments apply in the case where the weighting is done by means of factor- and regression analysis. The latter should also take care of the consistency of timing at peaks and troughs. As far as economic significance is concerned, important sectors of the economy are not covered by the survey (amongst others agriculture, mining and banking) but because of the high degree of economic interdependence between the manufacturing sector, the wholesale, retail and motor trade and the sectors which are not surveyed, it may be argued that the subjective opinions expressed in the survey are representative of a much wider field of economic activity.

The high intercorrelations between the different balance series necessitate a reduction in the dimensions of the explanatory data base before regression relationships between the balance series and a quantitative business cycle indicator can be explored. The original data base is reduced to three factors by means of principal components analysis followed by a Varimax factor rotation in order to provide a representation of the factor loadings matrix which facilitates interpretation. This rotation simply transforms the axes retained from the principal components solution into a new orthogonal space of the same dimensions. The procedure is aimed at improving the ease of interpretation of the reference
dimensions. The rotated factor matrix is presented in Table 1. It is based on observations over the period 1974:4–1984:4.

The total variance accounted for is 77.6%. Unfortunately the object of obtaining a relatively high loading on a single factor only is not achieved which complicates the identification of the underlying factors. The inverse relationship between the stock variables and the rest of the pro-cyclical variables is emphasized by the above factor loadings.

In the factor analysis standardized values of the survey variables are used, therefore the original balance series are standardized over the period 1974:4–1984:4 before the three factors, F1, F2 and F3, are calculated as linear combinations of the survey variables. CI, van Coller’s coinciding indicator averaged on a quarterly basis and transformed to percentual yearly changes on a quarterly basis.

\[ CIC_t = (CI_t - CI_{t-4})/CI_{t-4} \times 100 \]

is used as the dependent variable in the regression analysis. The rationale for using percentual changes in the case of variables measured relative to the same quarter in the previous year is to be found in the work of Anderson (1952)

CI, 4 / CI, 4 x 100

Meinhold & Singpurwalla (1983) ‘demystified’ the Kalman filter by putting it in the context of Bayesian inference. Their exposition was as follows.

Given a data set \( Y_t, Y_{t-1}, \ldots, Y_1 \) where \( Y_t \) is dependent on a non-observable quantity \( \theta_t \) known as the state of nature and the object is inference about \( \theta_t \), the observation equation describes the relationship between \( Y_t \) and \( \theta_t \) as

\[ Y_t = F_t \theta_t + v_t. \]

\( \theta_t \) is a scalar or vector independent of the dimension of \( Y_t \), and \( v_t \) is the observational error which is assumed to be normally distributed with mean 0 and variance \( V_t \).

The state of nature changes over time according to the systems equation:

\[ \theta_t = G_t \theta_{t-1} + w_t. \]

The systems error is \( w_t \) which is also assumed to be normally distributed with mean 0 and variance \( W_t \). Furthermore it is assumed that \( v_t \) and \( w_t \) are independent.

The Kalman filter is a recursive method of inference about \( \theta_t \) based on a Bayesian procedure.

At time \( t-1 \) the knowledge about \( \theta_{t-1} \) conditional on the information \( Y_{t-1} \) is embodied in the following probability statement

\[ (\theta_{t-1}/Y_{t-1}) \sim N(\tilde{\theta}_{t-1}, \Sigma_{t-1}) \]

where \( \tilde{\theta}_0 \) and \( \Sigma_0 \) are best guesses for the mean and variance of \((\theta_{t-1}/Y_{t-1})\) at time 0. Looking ahead at \( t \) in two stages, the best choice of \( \theta_t \) before \( Y_t \) is observed is given by the systems equation, therefore

\[ (\theta_t/Y_{t-1}) \sim N(G_t \tilde{\theta}_{t-1}, R_t = G_t \Sigma_{t-1} G_t' + W_t). \]

This is the prior distribution of \((\theta_t/Y_{t-1})\). Define \( e_t \) as the error made in the prediction of \( Y_t \) from time \( t - 1 \), i.e.

\[ e_t = Y_t - \bar{Y}_t = Y_t - F_t G_t \tilde{\theta}_{t-1}. \]

Observing \( Y_t \) is thus equivalent to the observation of \( e_t \) and after \( Y_t \) has been observed

\[ P(\theta_t/Y_{t-1}, Y_t) = P(\theta_t/e_t, Y_{t-1}). \]

It can now be shown that the posterior distribution of \( \theta_t \) is given by

\[ (\theta_t/e_t, Y_{t-1}) \sim N(G_t \tilde{\theta}_{t-1} + R_t F_t' (V_t + F_t R_t F_t')^{-1} e_t, \]

\[ R_t - R_t F_t' (V_t + F_t R_t F_t')^{-1} F_t R_t. \]
The procedure may be summarized in the following way:

(i) After time \( t = 1 \) the posterior distribution of \( \theta_{t-1} \) has mean \( \hat{\theta}_{t-1} \) and variance \( \Sigma_{t-1} \).

(ii) At time \( t \), form a prior distribution with mean \( G_{t} \hat{\theta}_{t-1} \) and variance \( R_{t} = G_{t} \Sigma_{t-1} G_{t}' + W_{t} \).

(iii) Evaluate a likelihood given \( e_{t} = Y_{t} - F_{t} \hat{\theta}_{t-1} \).

(iv) This leads to a posterior distribution of \( \theta_{t} \) with mean \( \hat{\theta}_{t} = G_{t} \hat{\theta}_{t-1} + R_{t} F_{t}' (V_{t} + F_{t} R_{t} F_{t}')^{-1} e_{t} \) and variance \( \Sigma_{t} = R_{t} - R_{t} F_{t}' (V_{t} + F_{t} R_{t} F_{t}')^{-1} F_{t} R_{t} \).

(v) Continue with the recursive procedure.

(vi) \( K_{t} = R_{t} F_{t}' (V_{t} + F_{t} R_{t} F_{t}')^{-1} \) is known as the Kalman filter.

(vii) The mean of the posterior distribution is the regression of \( \theta_{t} \) on \( e_{t} \).

(viii) The mean of the posterior distribution is therefore equal to the sum of the mean of the prior distribution and the regression coefficient of \( \theta_{t} \) on \( e_{t} \), multiplied by the error committed in predicting the next observation.

The aforementioned principles can now be applied to a dynamic linear system consisting of \( P \) observation equations and \( q \) systems equations:

\[
Y_{t} = H_{t} S_{t} + U_{Yt} + v_{t}
\]

\[
P \times 1 = p \times q \times q \times 1 \quad p \times 1
\]

\[
S_{t} = \Phi_{t} S_{t-1} + U_{t} + w_{t}
\]

\[
q \times 1 = q \times q \times q \times 1 \quad q \times 1 \quad q \times 1
\]

The usual assumptions apply. Given \( Y_{1}, Y_{2}, \ldots, Y_{t-1} \), the problem is to estimate \( S_{t} \). If \( k = t - 1 \), the problem is referred to as Kalman filtering, if \( k > t - 1 \) it is referred to as Kalman forecasting, and if \( k < t - 1 \) it is a problem in Kalman smoothing.

It can be shown that the minimum mean squared estimate of \( S_{t} \) based on \( Y_{1}, Y_{2}, \ldots, Y_{t-1} \), can be recursively expressed as

\[
\hat{S}_{q,t-1} = U_{Yt} + \Phi_{t} \hat{S}_{q,t-1}/t-2 + K_{q,t-1} (Y_{t-1} - U_{Yt-1} - H_{t-1} \hat{S}_{q,t-2})
\]

where \( K_{q,t-1} \) is called the Kalman gain for one stage prediction and is given by

\[
K_{q,t-1} = \Phi_{t} \Sigma_{q,t-1}/t-2 H_{t-1} (H_{t-1} \Sigma_{q,t-1}/t-2 H_{t-1} + \Sigma_{q,t-2})^{-1}
\]

and

\[
\Sigma_{q,t-1}/t-2 = E(\hat{S}_{q,t-1}/t-2 - \hat{S}_{q,t-2})
\]

\[
\Sigma_{q,t-2} = (\Phi_{t} - K_{q,t-1}/t-1 H_{t-1}) \Sigma_{q,t-1}/t-2 + \Sigma_{q,t-2}
\]

The estimate \( \hat{S}_{q,t-1} \) is obtained by predicting forward the estimate of the immediately preceding period and then correcting with a weighted combination of the components of the innovation vector \( Y_{t-1} - U_{Yt-1} - H_{t-1} \hat{S}_{q,t-2} \).

The filtered estimate of \( S_{t} \) based on \( Y_{1}, Y_{2}, \ldots, Y_{t} \), can be recursively expressed as

\[
\hat{S}_{q,t} = U_{Yt} + \Phi_{t} \hat{S}_{q,t-1}/t-1 + K_{t} (Y_{t} - U_{Yt} - H_{t} (U_{Yt} + \Phi_{t} \hat{S}_{q,t-1}/t-1))
\]

where \( K_{t} \) is the Kalman filter.

\[
K_{t} = \Phi_{t} H_{t}^{-1} K_{t+1}/t
\]

\[
\Sigma_{q,t} = (I - K_{t} H_{t}) \Sigma_{q,t-1}/t
\]

The results

In the analysis of the results presented here, it should be kept in mind that not all the forecasts are made at the same point in time. The following diagram presents the earliest stage at which it is possible to apply a particular technique (time is measured in months and the aim is to forecast the value of \( CI_{t} \) for the quarter ranging from \( t - 3 \) to \( t \)):

\[
\begin{array}{cccc}
A & & B \\
\hline
t-3 & t-2 & t & t+1 & t+3
\end{array}
\]

At point A, two weeks into the quarter, expectations data become available and the Smit/Kotze technique can be applied and an expectation for \( CI_{t} \), namely \( CI_{t}^{(1)} \), can be calculated. At the same time the Kalman filter estimate for an unknown \( S_{t-1} \) can be calculated. At point B, one month into the next quarter, quantitative information about the lagged variables in the Barr model becomes available and the quantitative model can be applied. At the same time the Kalman filter estimate for a known \( S_{t-1} \) can be calculated.

Model 1

In this model a weak generator of \( CI_{t} \) is tested. The system equation contains van Coller’s leading indicator as an explanatory variable, whilst it is well known that the lead time of this variable tends to be highly erratic. Expectations are generated using an adaptive expectations mechanism.

\[
CI_{t} = 6.61846 + 0.06218 LI_{t-1} + 0.86465 CI_{t-1}
\]

\[
\begin{array}{ccc}
(1.20) & (2.68) & (14.86)
\end{array}
\]

\[
R^{2} = 0.901; \quad SEE = 2.03654; \quad PF: 1975.2 - 1984.3
\]

\[
CI_{t}^{(1)} = 7.87296 + 0.60393 CI_{t-1}^{(1)} + 0.31869 CI_{t-1}
\]

\[
\begin{array}{ccc}
(1.11) & (4.79) & (2.73)
\end{array}
\]

\[
R^{2} = 0.840; \quad SEE = 2.38950; \quad PF: 1975.2 - 1984.3
\]
\[ \Sigma_{1} = (2.38950)^{2} = 5.7097 \]
\[ \Sigma_{2} = (2.03654)^{2} = 4.1475 \]
\[ \Sigma_{1/0} = 1.000 \]

The variables CI, and \( \Sigma_{1} \) have been defined before and LI represents van Coller's leading indicator. All error measures reported in Tables 2, 3 and 4 are over the full sample period.

In both cases (\( S_{-1} \) unknown and \( S_{-1} \) known) the filtered estimates perform better than the estimates based on the systems equation only. Both of these models, however, are inferior to the Smit/Kotze expectations generating mechanism. This is due to the poor specification of the systems equation.

**Table 2** Root mean square errors: Model 1

<table>
<thead>
<tr>
<th>Smit/Kotze</th>
<th>( S_{-1} ) unknown</th>
<th>( S_{-1} ) known</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5.0190 (Systems equation)</td>
<td>2.002 (Systems equation)</td>
</tr>
<tr>
<td></td>
<td>4.2011 (Filtered)</td>
<td>1.872 (Filtered)</td>
</tr>
</tbody>
</table>

Model 2

In this model a better specification of the systems equation is used, using the variables proposed by Barr in a Koyck lag scheme. The same expectations generating mechanism as used in Model I applies.

\[ \text{CI}_{t} = 32.38065 + 0.2105 \text{MV}_{t-1} + 0.00142 \text{DEMPOS}_{t-1} + 1.41889 \text{IMPORT}_{t-1} - 0.0094 \text{LOAND}_{t-1} \]
\[ R^2 = 0.966; \; \text{SEE} = 1.29266; \; \text{PF:} 1975:2-1984:3 \]

\[ \text{CI}_{t}^{*} = 7.90723 + 0.21530 \text{CI}_{t-1} + 0.70795 \text{CI}_{t-1}^{*} \]
\[ R^2 = 0.87; \; \text{SEE} = 1.85962; \; \text{PF:} 1975:2-1984:4 \]

\[ \Sigma_{t} = 1.67676^2 = 2.8115 \]
\[ \Sigma_{t}^{*} = 1.85962^2 = 3.4582 \]
\[ \Sigma_{1/0} = 1.000. \]

All the variables have been defined previously.

**Table 3** Root mean square errors: Model 3

<table>
<thead>
<tr>
<th>Smit/Kotze</th>
<th>( S_{-1} ) unknown</th>
<th>( S_{-1} ) known</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.1572 (Systems equation)</td>
<td>1.6208 (Systems equation)</td>
</tr>
<tr>
<td></td>
<td>2.1433 (Filtered)</td>
<td>1.6110 (Filtered)</td>
</tr>
</tbody>
</table>

Once more it is evident that the Kalman filter improves the accuracy of the estimate of \( \text{CI}_t \).

**Conclusion**

This article has explored an approach aimed at using qualitative expectations data as well as the lead time in the publication of this data. Three approaches have been compared: an approach using quantitative data, one based on qualitative data and one integrating the two data sets via the Kalman filter. For three different models the Kalman filter has led to improved forecasts and further research into the application of the Kalman filter to qualitative data seems a worthwhile pursuit.

**Notes**

1. Models 1 and 2 do not use the same factor loadings as presented in Table 1. This is due to the fact that the principal components analysis has been done twice over the past ten quarters. The factors loadings have remained very stable over time and the addition of new information.
2. This breakdown was already evident in Model 2 where the LOAND variable switched signs in the systems equation.
References


