On the construction of the european sentiment indicator

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

On the Construction of the European Economic Sentiment Indicator*

Sarah Gelper† and Christophe Croux‡

†*Erasmus School of Economics, Burgemeester Oudlaan 50, 3062 Rotterdam, The Netherlands (e-mail: gelper@ese.eur.nl)*
‡*Faculty of Business and Economics, Katholieke Universiteit Leuven, Leuven, Belgium (e-mail: christophe.croux@econ.kuleuven.be)*

Abstract

Economic sentiment surveys are carried out by all European Union member states and are often seen as early indicators for future economic developments. Based on these surveys, the European Commission constructs an aggregate European Economic Sentiment Indicator (ESI). This paper compares the ESI with more sophisticated aggregation schemes based on statistical methods: dynamic factor analysis and partial least squares. The indicator based on partial least squares clearly outperforms the other two indicators in terms of comovement with economic activity. In terms of forecast ability, the ESI, constructed in a rather *ad hoc* way, can compete with the other indicators.

I. Introduction

Every month, the European Commission publishes the European Economic Sentiment Indicator (ESI). The ESI is a survey-based indicator that aims to get insights into the beliefs of economic agents, both from the demand and supply sides of the economy. If consumers and manufacturers feel confident about the current and future economic situations, they might increase their consumption and production respectively. Moreover, the sentiment data provide new information as they are available earlier than most economic indicators like GDP or industrial production. These reasons, together with the growing integration of the European market, motivated

us to compare the forecasting performance of the ESI with forecasts based on alternative methods for weighting the underlying survey data.

The data for constructing an EU aggregate sentiment indicator are based on surveys carried out in all member states of the European Union. There are four business surveys, for the industrial, service, construction and retail sectors, and one consumer survey. For each country, a subset of 15 questions from these surveys is used to construct the ESI, resulting in a large number of sentiment series. The weights of the components are based on intuitive economic reasoning (more details are given in section II). The research question in this paper is whether the construction of the ESI can be improved. In other words, we investigate whether other aggregation methods, using the same sentiment components, may result in more informative indicators. In particular, we compare the ESI with sentiment indicators obtained by data-driven aggregation methods, namely the dynamic factor (DF) model as in Stock and Watson (2002), and the partial least squares (PLS) approach. We compare these indicators with the ESI in four respects: (i) the evolution of the indicators over time; (ii) the importance given in the aggregation scheme to each of the European countries and to each of the five surveys; (iii) the comovement of the economic indicators with economic activity; and (iv) their predictive power for industrial production growth both at the national and EU aggregate levels.

The predictive power of sentiment surveys is addressed in numerous studies. Hansson, Jansson and Lof (2005) study the forecasting performance of business survey data in Sweden. They use a DF model and find good results for forecasting GDP growth. A related study focussing on sentiment indicators is Slacalek (2005). He applies a DF model to the components of the Michigan sentiment survey. The resulting factors are found to be a stable predictor of US consumption growth. Our study differs from Slacalek (2005) as we explicitly compare different aggregation schemes and their out-of-sample forecasting performance. Furthermore, we do not limit our attention to consumer sentiment but combine this with results from production surveys, and work in a European context.

The predictive power of national indicators results in mixed findings that strongly depend on whether an in-sample or out-of-sample testing framework is used (e.g. Lemmens, Croux and DeKimpe, 2005, and references therein). A recent article by Cotsomitis and Kwan (2006) finds that the out-of-sample evidence for the forecasting power of national ESI and consumer confidence indicators for household spending is very limited. The research question addressed by this paper is to find out whether it is possible to improve the forecasting performance of the European ESI if different aggregation schemes are used. It will turn out that the indicator based on PLS is superior to the ESI in two respects. It shows stronger comovement with economic activity and it achieves better forecast performance at longer horizons.

The remainder of this paper is organized as follows. Section II first clarifies how the European Commission constructs the ESI and then explains the indicators based on the factor model and on the PLS method. A detailed comparison of these three

indicators can be found in section III. The comovement between ESIs and economic activity is studied in section IV. Section V outlines the framework to test for the predictive power and compares the forecast performance of the aggregate sentiment indicators. Finally, section VI concludes.

II. Constructing a European aggregate sentiment indicator

The purpose of constructing an aggregate indicator is to summarize the information contained in a large number of series into one single indicator series. In our setting, a series corresponds to a particular question from one of the five sentiment surveys, and we have such a series for every EU member state. In total, 15 sentiment components are retrieved from the five different surveys. Each component corresponds to a survey question, and is expressed in *balance*, i.e. the percentage of positive answers minus the percentage of negative answers to that question for a particular country. A study by Driver and Urga (2004) confirms that the balance is an appropriate way to summarize all individual answers to a certain survey question. We use surveys from 15 European countries, the member states of the EU in 1995, resulting in a total of 225 time series. Our aim is to find a method to summarize these 225 series in one indicator which can be interpreted as reflecting general European economic sentiment. Three aggregation schemes are considered: the methodology used by the European Commission to construct the ESI, the DF model and the PLS method. These three methods have in common that they construct a linear combination of the original sentiment series, but they differ in the way the weights for the linear combination are calculated.

The first aggregation method, which is used by the European Commission to construct the ESI, proceeds in two steps. In a first step, each component is aggregated over the member states using specific country weights. Then, in a second step, these 15 component series are aggregated by making use of survey weights to end up with one single indicator. For a component j, the weight of country i for month t is denoted by $w_{i,j,t}$, and given as a 2-year moving average by

$$w_{i,j,t} = \frac{v_{i,j,t} + v_{i,j,t-12}}{2}, \quad \text{with } v_{i,j,t} = \frac{X_{i,j,t}}{X_{\text{EU},j,t}}. \tag{1}$$

Here, $X_{i,j,t}$ is a certain economic variable measured for member state i in year t and $X_{\text{EU},j,t}$ is the European equivalent. The economic variable $X_{i,j,t}$ differs according to the survey from which component j originates. For the industrial, construction and services sentiment surveys, $X_{i,j,t}$ is the yearly gross value added at constant prices in the respective sector for country i at time t. For retail and consumer sentiments, $X_{i,j,t}$ represents the yearly private final consumption expenditure at constant prices for country i at time t. Equation (1) allows the country weights to be time varying. In practice, they only vary slightly over the years. A weighted sum over all countries yields the value of the sentiment components at the EU level at time t.

After the 15 EU-level sentiment components have been obtained, these are aggregated using survey weights. These weights are based on two criteria.[1] First, they should reflect the importance of the corresponding sector in the total economy. For instance, the service sector is responsible for a larger amount of total GDP than the retail sector, and therefore gets a larger weight. Secondly, the more the survey results from a certain sector comove with GDP, the more weight this survey should get. Taking these two criteria into account, the European Commission decided on the following weights for the five surveys: 40% for industrial confidence, 30% for services, 5% for retail trade, 5% for construction and 20% for consumer confidence. The weight of each survey is then equally divided over the different questions within the survey.

We consider two competing methods for constructing an aggregate sentiment indicator. Both methods are based on the same hypothesis, namely that there is one underlying driving factor influencing all observed sentiment components $x_{k,t}$, where $k = 1, \ldots, 225$. Every predictor series is written as a combination of a common component F_t, called the underlying factor, and an idiosyncratic component $e_{k,t}$:

$$x_{k,t} = \lambda_k F_t + e_{k,t}, \qquad (2)$$

where λ_k is the factor loading of component k. In general, there can be more than one underlying factor. Here, we restrict ourselves to only one factor, which is in accordance with the assumption that there is one single driving force influencing all economic sentiment components in all countries. The factor F_t is considered as an aggregate sentiment indicator. We consider two ways of estimating or identifying the common component F_t, the DF method and the method of PLS.

The methodology of the DF model, proposed by Stock and Watson (2002), makes use of the method of principal components. More specifically, the underlying factor F_t in model (2) is estimated by extracting the first principal component from the full set of 225 sentiment components. The first principal component derived from a large data set is the linear combination of the individual series that maximizes the variance of the factor, subject to the constraint that the sum of all squared weights equals one (for a detailed discussion, see Stock and Watson, 2002). As opposed to the weights for the ESI, the weights in the factor model are exclusively based on the past values of observed sentiment component series. In a forecasting context, Banerjee, Marcellino and Masten (2005) and Marcellino, Stock and Watson (2003) find that the use of a DF model constructed from many economic indicators (but not including sentiment indicators) improves forecasts of aggregate European real economic variables when compared with univariate modelling.

As another competing method for reducing the dimension of the 225 series into one single series, we consider PLS. While neither the factor model nor the ESI construction methods take the variable to be predicted into account, the method of PLS does. Hence, PLS will construct another indicator for every economic variable to be

predicted. The weights for constructing the common component are chosen such that the covariance between the aggregate indicator and the variable to predict is maximized. As such, the resulting indicator takes into account the covariance with the variable to predict, hereby aiming at a better forecasting performance. The PLS method first standardizes all series. Then a simple recursive computing scheme yields a sequence of underlying factors. In our case, when only one common factor needs to be constructed, the weights given by the PLS method are simply the covariances between the variable to predict and the predictor variables. The computational effort required for a PLS analysis is negligible, which is one reason for its popularity. Although PLS has successfully been applied in chemistry and engineering, it is much less known in economics. Reviews of the use of PLS, with additional references, can be found in Wold (2006) and Helland (2006).

III. Comparison of the sentiment indicators

In a first stage of our empirical analysis, we compare the evolution of the three proposed monthly indicators: ESI_t as constructed by the European Commission, DF_t obtained from DF analysis and PLS_t resulting from the PLS method. We use the survey data for all 15 EU member states before the enlargement in 2004 for the time range of April 1995 to November 2005. In principle, this results in a total of 225 sentiment series (as discussed in section II). However, as a number of surveys started at a later date, our data set only comprises 160 complete series. We prefer to work with the complete time series only: if entire time sequences are missing at the beginning of a series, no standard imputation procedures are available. These 160 series are used to construct common indicators based on factor analysis and PLS. Recall that PLS_t depends on the series to predict, and in this section the PLS indicator is aimed at predicting industrial production growth at the European level.

Figure 1 shows the evolution of the indicators from September 1998 to November 2005.[2] The DF and PLS methods are applied to the component series in differences, as they are (borderline) non-stationary in levels. The stationarity condition is required for consistency of the DF model (see Stock and Watson, 2002). Furthermore, the differenced sentiment series are at least as informative as the series in levels, as the most important information is to know whether the general sentiment increases or decreases. Nevertheless, the sentiment indicator in levels provides a more appealing graphical representation and can easily be reconstructed from the differences. The DF and PLS sentiment indicators are recursively obtained: at each time point t, the indicator is extracted from all sentiment component series up to moment t. Only information from the past is included in the calculation of the current weights, so that the indicators are computed in real time. It follows from this updating

---

[2]The first 40 months are used as a start-up period long enough to compute a reliable indicator from it.
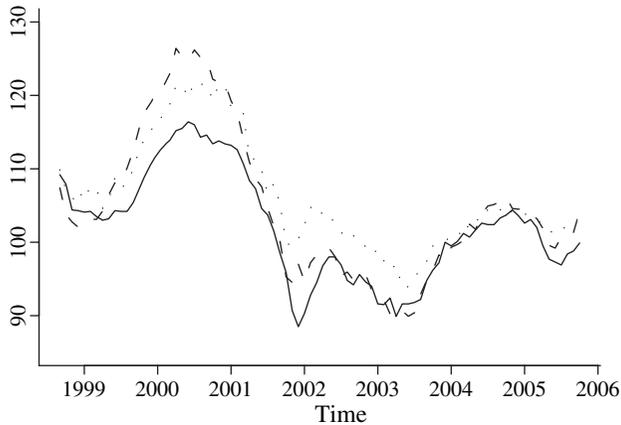
Figure 1. Three sentiment indicators: ESI$_t$ (solid line), DF$_t$ (dashed line) and PLS$_t$ (dotted line), from September 1998 to November 2005

procedure that the DF and PLS weights of the sentiment component series are time varying.

Figure 1 shows that the three indicators move closely together. Although the construction of the ESI is not based on formal statistical arguments like the other two indicators, the latter have an instantaneous correlation with the ESI of about 0.95. Important cross-correlations at different leads and lags are also present. One should, however, bear in mind that all series in Figure 1 are close to being unit-root processes, which may result in spurious correlations.[3] In fact, the Johansen test reveals that the three series are cointegrated ($P = 0.01$). The instantaneous correlation between the differenced series equals 0.48 between ESI$_t$ and DF$_t$, and 0.38 between ESI$_t$ and PLS$_t$, which are still fairly high values.

Apart from comparing the sentiment indicators as such, it is also interesting to compare the weights of the component series assigned by the three methods. By adding up the weights of the 15 questions for each country, we obtain country weights. Similarly, summing the weights over the countries for each survey yields the survey weights. Figure 2 plots the country weights of the DF approach (panel a) and the weights of the PLS approach (panel b) versus the ESI weights. An ordinary least squares regression line is added. In both panels, the slope of the regression line is significantly positive. The statistically based selection of country weights, resulting from factor extraction by DF and PLS, is in line with the more intuitive and economic arguments for the ESI weights. For example, Germany is a large country and has always been considered as an important member of the European Union. Accordingly, Germany receives the highest weight in the construction of the ESI using the PLS approach, while the DF analysis attributes the third largest weight to Germany. Similarly, France receives high weights in all 3 constructed indicators. On the other hand, Belgium, for instance, is much smaller and gets a low weight in the

---

[3]Augmented Dickey–Fuller tests yield *P*-values of 0.42 for the ESI, 0.17 for DF and 0.49 for PLS.
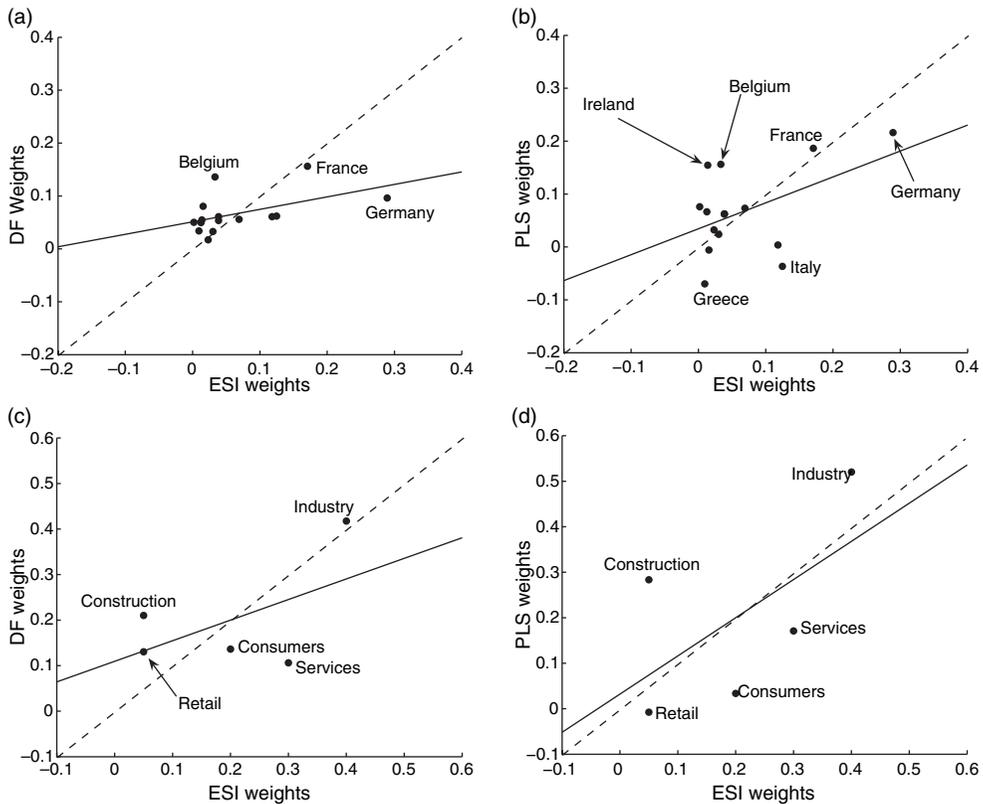
Figure 2. Scatter plots of the country weights as obtained by dynamic factor analysis (DF, panel a) and partial least squares (PLS, panel b) versus ESI together with scatter plots of the survey weights as obtained by dynamic factor analysis (DF, panel c) and partial least squares (PLS, panel d) versus ESI. An ordinary least squares regression fit (solid line), together with the 45° line (dashed line), is added (a) Country weights; (b) Country weights; (c) Survey weights; (d) Survey weights

construction of the ESI. It is, however, an important country according to the DF and PLS methods. The likely reason is that Belgium is a very open economy, with high exports to neighbouring countries. Hence, the outcomes from Belgian surveys are more informative than one would expect from its country size. The PLS method also attaches much more weight to Ireland than the ESI does. On the other hand, PLS gives much less weight to Greece and Italy than the ESI; in fact these weights are even slightly negative. This suggests an atypical behaviour of these countries when it comes down to predicting European level industrial production growth. Panels (c) and (d) in Figure 2 compare the survey weights. We see a positive correlation between the weights used for the ESI and the weights from both DF and PLS. In particular, the industrial sector receives a high weight, as it represents a large percentage of total European GDP. For the PLS method this was to be expected, as the PLS method is choosing the weights using information on the covariance with the growth of industrial production at the European level.

IV. The comovement of economic activity and sentiment

The main purpose of constructing an aggregate sentiment indicator is to get an early indication of future economic fluctuations. This suggests that there should be some degree of comovement between economic activity and sentiment. To quantify this comovement, we follow the approach suggested by den Haan (2000) who makes use of vector autoregressive (VAR) models. Let z_t be the bivariate time series containing industrial production growth IP_t, as a measure for economic activity, and an ESI S_t. The VAR model for direct forecasting at horizon h is given by

$$z_{t+h} = \alpha + \phi_0 z_t + \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + \varepsilon_t, \tag{3}$$

where ε_t is a bivariate vector of serially uncorrelated innovations with mean zero. Instantaneous cross-correlation between the two components of ε_t is allowed. It is argued in den Haan (2000) that if the out-of-sample forecast errors of the VAR model are cross-correlated, then there is comovement between the two components of z_t. In our setting, this means that the ESI S_t comoves with economic activity. The sentiment indicator S_t is either the ESI in differences (denoted by ESI_t), or the indicator derived from factor analysis (DF_t) or from partial least squares (PLS_t), where the latter two are constructed based on the differenced sentiment components.

Model (3) is fitted by ordinary least squares and the lag length p is optimally chosen according to the Bayesian information criterion. From the fitted model, a bivariate series of h-step-ahead forecast errors is obtained and the correlation between them is computed. To test whether an observed correlation is significant or not, we use the residual bootstrap approach suggested by den Haan (2000). A significant positive correlation can be interpreted as follows: accounting for the past of both economic sentiment and industrial production growth, an over-prediction of economic sentiment coincides with an over-prediction of industrial production growth. In other words, economic sentiment and industrial production growth show similar deviations from what could be expected according to the VAR model.

The approach of den Haan (2000) has recently been applied in the context of economic sentiment by Taylor and McNabb (2007). Our study is different in at least four respects. First, Taylor and McNabb (2007) consider only four countries (UK, France, Italy and the Netherlands), while we have a larger set of countries (we add Belgium, Germany, Luxembourg and Denmark). Moreover, we do not focus only on the country-specific level but also include the EU aggregate. Secondly, Taylor and McNabb (2007) study consumer and business confidence separately. Our aim is to study the construction of a general ESI which includes both the demand and supply sides of the economy. Thirdly, while Taylor and McNabb (2007) study GDP data, we use industrial production as an approximation for economic activity (as in den Haan, 2000). This allows us to work with monthly instead of quarterly data. Finally, the VAR model fitted in Taylor and McNabb (2007) is based on the original series in levels, while we use the growth rate of industrial production and first differ-

ences of the sentiment indicator. These time series are stationary, avoiding spurious results.

Figure 3 shows the correlation structures between the forecast errors for IP_t and S_t for $h = 1, 2, \ldots, 12$. The correlations are computed for IP_t equal to the monthly industrial production growth in each of the eight countries under consideration, and for the EU. The three panels of Figure 3 show the correlation results for the economic sentiment indicators ESI_t, DF_t and PLS_t respectively. Correlations significant at the 5% level according to the bootstrap procedure are indicated by a bold dot. Figure 3 provides evidence that the three sentiment indicators correlate most strongly with the EU aggregate IP_t. This suggests that an aggregate sentiment indicator primarily comoves with EU aggregate economic activity, rather than with the country-specific counterpart. Moreover, the PLS_t has a noteworthy higher correlation with the EU aggregate IP_t (correlations fluctuating around 0.8 for different horizons) than the other two sentiment indicators (correlations of about 0.4). The indicator obtained by PLS thus shows a much stronger comovement with EU aggregate economic activity than the other two sentiment indicators, both for short and longer horizons.

Furthermore, there are some interesting patterns for the country-specific forecasts. First of all, for the ESI_t, we only find significant correlations with the German IP_t. This may result from the high weight accorded to Germany in constructing the ESI, as can be seen from Figure 2. Secondly, the DF_t has significant correlations with the IP_t of both Italy and Belgium. The latter is consistent with Belgium's high weight in the construction of the DF_t. Finally, and most importantly, the PLS_t correlates significantly with all country-specific IP series, except for Luxembourg, and is thus clearly the best comoving indicator. We conclude that the PLS_t is the indicator that most strongly comoves with economic activity, at the EU aggregate as well as at the country-specific level.

V. Forecasting using sentiment indicators

This section compares the ESI with the indicators obtained by DF analysis and the PLS method in terms of forecast performance. In particular, we build and evaluate point forecasts of industrial production growth using the three different sentiment indicators. One might argue that only industry-survey data should be used for forecasting industrial production. However, we are mainly interested in the forecasting power of a general sentiment indicator. Moreover, Lee and Shields (2000) provide empirical evidence for sentiment–output interaction between different sectors.

In our context, the question of forecasting power is twofold. In a first stage, we investigate the benefit of constructing a EU aggregate sentiment indicator for forecasting purposes. More specifically, we study whether any of the three indicators has predictive power. In a second stage, we study the benefit of using a statistical procedure like DF modelling or PLS instead of using the existing ESI.

The model we study is very simple and given by

$$IP_{t+h} = \alpha + \beta S_{t-1} + \varepsilon_t^1 \tag{4}$$
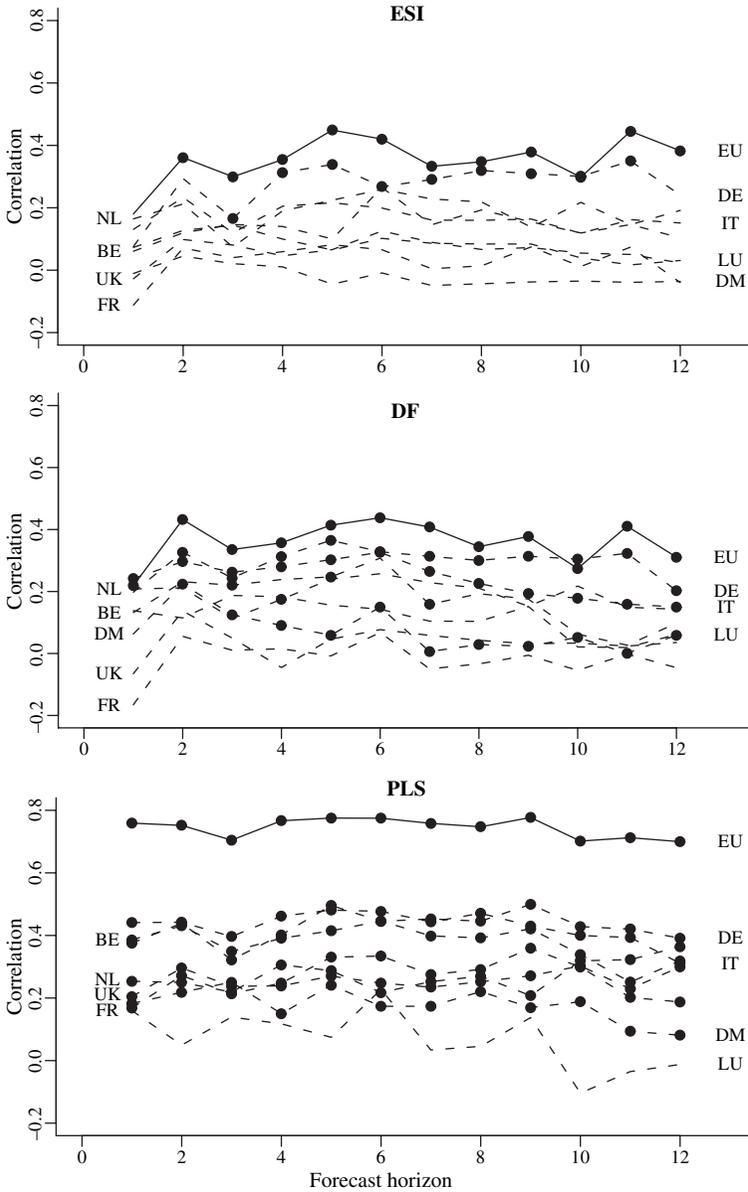
Figure 3. Correlations of the forecast errors using the bivariate VAR model (3) for horizons 1–12 for three sentiment indicators (ESI_t in the top panel, DF_t in the middle panel and PLS_t in the lower panel) with industrial production growth in eight European countries (dashed lines) and the EU aggregate industrial production growth (solid line). Significant correlations are indicated by a dot

where ε_t^1 is a noise component with mean zero and $h \geq 0$ denotes the forecast horizon. Recall from section IV that IP_t is the industrial production series in log differences, and that both IP_t and S_t are stationary. We use model (4) because of its simpli-

city. When $h = 0$, for example, a positive β means that an increase in the aggregate sentiment in the previous month indicates a boost of this month's industrial production growth. Put differently, only by observing the most recently available sentiment indicator S_{t-1} can we gain information about current and future economic activity. Sentiment indicators are indeed often used as a barometer of the economy. No additional lags nor other macroeconomic variables are included in equation (4), as the sentiment indicator serves as a simple indication readily accessible to the media and non-academics that can be followed without any further econometric analysis.

The first step of our analysis studies the usefulness of an aggregate EU sentiment indicator in terms of forecasting power. We compare model (4) with a benchmark model that assumes industrial production to be a random walk. In that case, S_{t-1} contains no information about current or future developments of IP_t and the appropriate model is given by

$$\text{IP}_{t+h} = \alpha + \varepsilon_t^0, \tag{5}$$

where ε_t^0 is zero-mean noise. Model (5) predicts the h-month-ahead industrial production growth as the average of all previously observed growth rates. If model (4) outperforms the benchmark model (5) for any sentiment indicator, an aggregate sentiment indicator is informative for current and future economic developments.

To test for the predictive power of the sentiment indicators, we work in an out-of-sample framework. As opposed to in-sample procedures, out-of-sample techniques test whether the 'real-time' forecast errors from a certain model are significantly smaller than those from the benchmark model. The 'real-time' forecast errors are constructed by a recursive forecasting scheme. The first R observations are used to forecast observation $R+h$, after which the forecast is compared with the realized value. This yields a first forecast error. Then, observation $R+1+h$ is forecasted using all observations up to period $R+1$ and again the forecast and the realized value are compared. This procedure continues until the end of the series and results in a series of h-step-ahead forecast errors. We selected $R = \lfloor T/2 \rfloor$, resulting in a sequence of $\lfloor T/2 \rfloor - h$ forecast errors. As the data range from April 1995 to November 2005, we have $T = 128$ monthly observations and $R = 64$. To compare the forecast performance of the four models under consideration, we compute their out-of-sample mean squared forecast error (MSFE), defined as the mean of all squared out-of-sample forecast errors obtained from the recursive scheme.

To remain in a pure out-of-sample framework, it is necessary to recompute the indicators obtained from the DF analysis or the PLS method in each step. More precisely, when forecasting the observation $R+s+h$, in step s of the recursive scheme, the weights for the 160 sentiment series are computed based on the first $R+s$ observations only. Out-of-sample tests require much computation time, but are conceptually more appealing as they mimic the process of true real-time predictions of future values. As the purpose of constructing a sentiment indicator is to have an early indication of the future status of economic variables, an out-of-sample testing framework is the most natural. There is, however, no general rule stating that the out-of-sample

testing is always preferred to in-sample procedures (see, for instance, Clements and Hendry, 2005; Inoue and Kilian, 2005).

To compare the models using a sentiment indicator as in equation (4) with the constant growth benchmark model (5), we perform a reality check for data snooping as proposed in White (2000) relying on the stationary bootstrap procedure developed by Politis and Romano (1994). The reality check tests for the existence of superior forecast performance of the best of a sequence of models with regard to the benchmark. The idea is that if one considers a large enough number of forecasting models and if only pairwise tests are used, one model is likely to be significantly better than the benchmark, merely by chance. In this first step of our forecast analysis, the benchmark is the constant growth model (5) and the three alternatives are given by model (4) where S_t is one of the three sentiment indicators ESI_t, DF_t or PLS_t. The null hypothesis states that the best among these three is no better than the benchmark and P-values are computed using the reality check.

The results are presented in Table 1. The four panels correspond to forecast horizon h, equal to 0, 1, 3 or 6 respectively. Note that for $h = 0$, we make a *nowcast* of current IP_t. Making nowcasts for macroeconomic variables is extensively studied in Giannone, Reichlin and Small (2008). For every forecast horizon, the MSFE ($\times 10,000$) for the constant expected growth model (5) is given for the different countries and the EU aggregate. The next columns present the relative MSFE values of model (4), where S_t is either of the three sentiment indicators, with respect to model (5). All relative MSFE values are smaller than 1, indicating that each indicator results in a smaller MSFE than the constant expected growth model, and this at every horizon and for every country. Moreover, the reported P-values for the White reality check in this first step are small. This indicates that the best of the three indicators has forecasting power superior to the naive random walk forecast model (5). We conclude that an EU aggregate sentiment indicator is informative for current and future developments of economic activity, both European and country-specific levels.

The White data-snooping check in 'step 1' compares nested models, which may lead to serious size distortions at finite samples. The reason is that one expects smaller values for the MSFE even if the null of equal forecast accuracy holds, as the benchmark model contains fewer parameters to estimate. White data snooping is not correcting for parameter estimation error, leading to size distortion and to a loss in power at finite samples (for a recent work on this topic, see Hubrich and West, 2008). Clark and West (2007) propose replacing the MSFE of the larger model by an adjusted MSFE, where this adjustment reduces the MSFE. The data-snooping check of White is based on bootstrapping the distribution of the maximum difference of the MSFE of a sequence of forecasting models with a benchmark model. In the nested case, one then simply needs to replace the MSFE of the larger model by the adjusted version. We implemented this approach and obtained, as expected, P-values smaller than or almost equal to the ones reported in Table 1. Making the finite-sample correction in 'step 1' did not change the conclusions of our analysis. Other approaches for forecast

TABLE 1

For every country and for the EU aggregate, mean-squared forecast error (MSFE, ×10,000)
for the constant growth forecast model (const) and relative MSFE with respect to the
constant growth model for the three sentiment indicators

	MSFE	Relative MSFE			P-value		MSFE	Relative MSFE			P-value	
	const	ESI_t	DF_t	PLS_t	Step1	Step2	const	ESI_t	DF_t	PLS_t	Step1	Step2
$h=0$							$h=1$					
BE	5.11	0.41	**0.40**	0.43	0.02	0.69	4.24	**0.55**	0.57	0.58	0.04	0.73
DM	20.20	0.35	0.34	**0.33**	0.00	0.05	17.37	**0.45**	0.53	0.55	0.00	0.98
DE	3.55	0.35	**0.34**	0.37	0.00	0.68	2.69	0.50	**0.48**	0.55	0.00	0.87
FR	2.12	0.39	0.40	**0.37**	0.00	0.87	1.60	**0.50**	0.51	0.63	0.01	0.97
IT	1.36	0.43	**0.42**	**0.42**	0.00	0.26	1.02	**0.57**	0.59	0.76	0.08	0.97
LU	28.29	**0.36**	**0.36**	0.37	0.00	0.76	18.99	0.55	0.55	**0.53**	0.02	0.27
NL	115.11	**0.34**	**0.34**	**0.34**	0.00	0.55	73.57	0.57	0.55	**0.54**	0.00	0.03
UK	2.83	0.38	0.38	**0.37**	0.07	0.69	2.31	**0.46**	0.47	0.51	0.05	0.98
EU	0.70	0.42	**0.41**	0.43	0.00	0.66	0.70	**0.50**	**0.50**	0.72	0.13	0.99
$h=3$							$h=6$					
BE	5.27	0.44	**0.40**	0.43	0.02	0.29	4.99	**0.43**	0.44	**0.43**	0.02	0.94
DM	15.86	**0.45**	**0.45**	**0.45**	0.00	0.61	15.09	0.47	0.47	**0.46**	0.00	0.30
DE	3.00	0.43	**0.41**	**0.41**	0.00	0.11	2.75	0.46	0.43	**0.42**	0.00	0.02
FR	1.66	0.51	0.49	**0.47**	0.00	0.08	1.48	0.57	0.56	**0.55**	0.00	0.38
IT	0.96	0.69	0.61	**0.60**	0.02	0.00	0.90	0.66	0.67	**0.63**	0.02	0.73
LU	19.77	0.57	0.57	**0.52**	0.03	0.63	15.85	0.65	0.64	**0.64**	0.01	0.15
NL	82.34	0.49	**0.48**	0.49	0.00	0.47	83.03	0.48	**0.47**	0.48	0.00	0.77
UK	2.06	0.52	**0.51**	**0.51**	0.03	0.17	1.92	0.56	0.57	**0.55**	0.04	0.58
EU	0.67	0.54	0.46	**0.45**	0.01	0.00	0.61	0.55	0.52	**0.48**	0.02	0.08

Notes: The best method is indicated in **bold**. Reality check *P*-values are given for the constant growth model as benchmark (Step 1) and for the model with ESI_t as benchmark (Step 2).

evaluation of pairs of nested models are possible. Rapach and Wohar (2006), for example, extended the residual bootstrap-based method of Kilian (1999).

Now that we have shown that constructing a EU aggregate sentiment indicator yields a useful barometer, the second step of our analysis investigates whether the indicators based on the DF model or the PLS method have more power than the ESI. Note that the benchmark prediction model (4), using the ESI, is not nested within the prediction models using the other indicators. Similarly, as in the first step of our analysis, we apply the White reality check for data snooping but now with the ESI_t as a benchmark. The null hypothesis we test for is that both sentiment indicators DF_t and PLS_t show no predictive power superior to the ESI_t. The corresponding *P*-values are presented in Table 1 (under the heading 'step 2'). In most cases, as Table 1 shows, there is no significant improvement in forecast accuracy compared with the ESI and the value of the relative MSFE is comparable for the three indicators. For $h=3$ and 6, however, the relative performance of PLS is the best. For 6-month-ahead forecasts, the PLS_t sentiment indicator even consistently has the smallest MSFE, although the difference with the ESI remains insignificant in most cases. We stress that the weights

used by the PLS do not only depend on the variable to be predicted but also on the forecast horizon. In this application, it turns out that revising the weights of the sentiment indicator according to the forecast horizon is worthwhile, and yields better forecast performance at longer forecast horizons. We conclude from Table 1 that in most cases the more sophisticated indicators do not have a significantly improved predictive power over the ESI. An important exception is the forecast of the EU aggregate at longer horizons, where the PLS indicator genuinely improves on the forecasting power of the ESI (at the 10% significance level). An analogous in-sample model comparison confirms our findings.

An alternative approach would be to study forecasting power within a Granger causality framework. This allows us to measure the *incremental* predictive power of the indicators with respect to the past of the series to be predicted. Testing for no Granger causality can be achieved by inserting lagged values of both industrial production growth and economic sentiment as predictors in model (4). Granger causality of sentiment indicators has been studied before by Carroll, Fuhrer and Wilcox (1994), Desroches and Gosselin (2002), Bryant and Macri (2005), Gelper, Lemmens and Croux (2007), among others, but most often only within an in-sample testing framework. We performed out-of-sample Granger causality tests, as a better proxy for a true forecast exercise. It turns out that a purely autoregressive model for IP_t performs well (confirming the results of Gelper *et al.*, 2007). No significant improvements in forecast performance were found when lagged values of the sentiment series were added. These series are thus informative indicators, but they do not contain additional forecast information beyond the past data on industrial production growth itself.

VI.   Conclusion

This paper compares the European ESI, as published monthly by the European Commission, with two other methods for constructing an aggregate European sentiment indicator. The two alternative ways of aggregating the 225 sentiment component series are based on statistical techniques: a DF analysis and the PLS method. The aggregation weights obtained by the different methods are related, although far from identical. The evolution over time of the different indicators, on the other hand, is very similar.

An aggregate sentiment indicator is interesting as it is expected to comove with economic activity and to provide early information for future economic developments. In terms of comovement, the indicator based on PLS is by far the best indicator, as we showed in section IV. This is hardly surprising because construction of the PLS indicator takes the variable to be predicted into account. The superior performance of the PLS indicator holds both at the national and EU aggregate levels. While the PLS method is currently rarely used in econometrics, the results obtained in this paper suggest that it should be included in the toolkit of an economic forecaster.

Besides studying comovement, we look at the more difficult problem of obtaining point forecasts of industrial production growth rates. It is shown that constructing

an aggregate EU sentiment indicator has significant out-of-sample predictive power for both national and EU aggregate industrial production growth rates. Although the European ESI seems to be constructed in a rather *ad hoc* way, its forecast performance is mostly comparable with that of other aggregation schemes that are based on statistical arguments. For longer forecast horizons, however, the PLS method outperforms the ESI. An advantage of the PLS method is that the weights used in the construction of the indicator do not only take the variable to be predicted into account but also the forecast horizon.

Finally, a Granger causality analysis showed that the sentiment indicators, either the ESI or the ones based on statistical criteria, do not have much additional explanatory power for industrial production compared with autoregressive forecasting methods. Sentiment indicators are useful as an early barometer of the economy, yielding a simple and accurate index that non-academics can follow but do not contain much relevant information that is not observable in the past of the industrial production series. It would be interesting to extend the present analysis to test the performance of the more sophisticated sentiment indicators for forecasting other quantities, like measures of output gaps used in monetary policy analysis.

Final Manuscript Received: August 2009

References

Banerjee, A., Marcellino, M. and Masten, I. (2005). 'Leading indicators for Euro-area inflation and GDP growth' *Oxford Bulletin of Economics and Statistics*, Vol. 67, pp. 785–813.

Bryant, W. and Macri, J. (2005). 'Does sentiment explain consumption?' *Journal of Economics and Finance*, Vol. 29, pp. 97–110.

Carroll, C., Fuhrer, J. and Wilcox, D. (1994). 'Does consumer sentiment forecast household spending? If so, why?' *American Economic Review,* Vol. 84, pp. 1397–1408.

Clark, T. and West, K. (2007). 'Approximately normal tests for equal predictive accuracy in nested models', *Journal of Econometrics*, Vol. 138, pp. 291–311.

Clements, M. and Hendry, D. (2005). 'Evaluating a model by forecast performance', *Oxford Bulletin of Economics and Statistics*, Vol. 67, pp. 931–956.

Cotsomitis, J. and Kwan, C. (2006). 'Can consumer confidence forecast household spending? Evidence from the European Commission business and consumer surveys', *Southern Economic Journal*, Vol. 72, pp. 597–610.

Desroches, B. and Gosselin, M. (2002). *The Usefulness of Consumer Confidence Indexes in the United States*, Bank of Canada Working Paper.

Driver, C. and Urga, G. (2004). 'Transforming qualitative survey data: performance comparisons for the UK', *Oxford Bulletin of Economics and Statistics*, Vol. 66, pp. 71–89.

Gelper, S., Lemmens, A. and Croux, C. (2007). 'Consumer sentiment and consumer spending: decomposing the Granger causal relationship in the time domain', *Applied Economics*, Vol. 39, pp. 1–11.

Giannone, D., Reichlin, L. and Small, D. (2008). 'Nowcasting: the real-time infor- mational content of macroeconomic data', *Journal of Monetary Economics*, Vol. 55, pp. 665–676.

den Haan, W. (2000). 'The comovement between output and prices', *Journal of Monetary Economics*, Vol. 46, pp. 3–30.

Hansson, J., Jansson, P. and Lof, M. (2005). 'Business survey data: do they help in forecasting GDP growth?' *International Journal of Forecasting*, Vol. 21, pp. 377–389.

Helland, I. (2006). 'Partial least squares regression', in Kotz S., Read B., Balakrishnan N. and Vidakovic B. (eds), *Encyclopedia of Statistical Sciences*, 16 Volume Set, 2nd edn, John Wiley & Sons, Inc., pp. 5957–5962.

Hubrich, K. and West, K. D. (2008). *Forecast Evaluation of Small Nested Model Sets*, Working Paper 1030, ECB.

Inoue, A. and Kilian, L. (2005). 'In-sample or out-of-sample tests of predictability: which one should we use?' *Econometric Reviews*, Vol. 23, pp. 371–402.

Kilian, L. (1999). 'Exchange rates and monetary fundamentals: what do we learn form long-horizon regressions?' *Journal of Applied Econometrics*, Vol. 14, pp. 491–510.

Lee, K. and Shields, K. (2000). 'Expectations formation and business cycle fluctuations: an empirical analysis of actual and expected output in UK manufacturing, 1975–1996', *Oxford Bulletin of Economics and Statistics*, Vol. 62, pp. 463–490.

Lemmens, A., Croux, C. and Dekimpe, M. (2005). 'On the predictive content of production surveys: a pan-European study', *International Journal of Forecasting*, Vol. 21, pp. 363–375.

Marcellino, M., Strock, J. and Watson, M. (2003). 'Macroeconomic forecasting in the Euro area: country specific versus area-wide information', *European Economic Review*, Vol. 47, pp. 1–18.

Politis, D. and Romano, P. (1994). 'The stationary bootstrap', *Journal of the American Statistical Association*, Vol. 89, pp. 1303–1313.

Rapach, D. and Wohar, M. (2006). 'In-sample vs. out-of-sample tests of stock return predictability in the context of data mining', *Journal of Empirical Finance*, Vol. 13, pp. 231–247.

Slacalek, J. (2005). *Analysis of Indexes of Consumer Sentiment*, Working Paper, German Insititute for Economic Research.

Stock, J. and Watson, M. (2002). 'Forecasting using principal components from a large number of predictors', *Journal of the American Statistical Association*, Vol. 97, pp. 1167–1179.

Taylor, K. and McNabb, R. (2007). 'Business cycles and the role of confidence: Evidence for Europe', *Oxford Bulletin for Economics and Statistics*, Vol. 69, pp. 185–208.

White, H. (2000). 'A reality check for data snooping', *Econometrica*, Vol. 68, pp. 1097–1126.

Wold, H. (2006). 'Partial least squares', in Kotz S., Read B., Balakrishnan N. and Vidakovic B. (eds), *Encyclopedia of Statistical Sciences*, 16 Volume Set, 2nd edn, John Wiley & Sons, Inc., pp. 5948–5957.