CENTRAL BANK OF CYPRUS
EUROSYSTEM

WORKING PAPER SERIES

Monitoring Forecasting Combinations with Semiparametric Regression Models

Antonis Michis

May 2012

Working Paper 2012-02

Central Bank of Cyprus Working Papers present work in progress by central bank staff and outside contributors. They are intended to stimulate discussion and critical comment. The opinions expressed in the papers do not necessarily reflect the views of the Central Bank of Cyprus or the Eurosystem.

Address
80 Kennedy Avenue
CY-1076 Nicosia, Cyprus

Postal Address
P. O. Box 25529
CY-1395 Nicosia, Cyprus

E-mail
publications@centralbank.gov.cy

Website
http://www.centralbank.gov.cy

Fax
+357 22 378153

Papers in the Working Paper Series may be downloaded from:
http://www.centralbank.gov.cy/nqcontent.cfm?a_id=5755

Monitoring Forecasting Combinations with Semiparametric Regression Models

Antonis Michis*

May 2012

Abstract

In this study, a modelling framework is proposed for evaluating the accuracy of forecasting combinations when the number of available forecasts is large and changes in time. Squared forecast errors are modelled with a semiparametric additive regression model where the linear part involves indicator variables reflecting the time period when the forecast is performed and the nonparametric part involves a smooth function of the number of individual forecasts entering the combinations. The partial regression estimates permit two-dimensional plots of the relationship between squared forecast errors and the number of forecasts entering the combinations and can be used to assess the contribution of additional forecasts in reducing the forecast errors. The method is demonstrated with six empirical applications using macroeconomic forecasts published by Her Majesty's Treasury.

Keywords: Forecasting combinations, semiparametric models, forecast errors.

JEL classification: C53, E27, E3.

* Central Bank of Cyprus. The author would like to thank Andrew Johnson of HM Treasury for valuable help in understanding the data. The opinions expressed in this paper are those of the author and do not necessarily reflect the views of the Central Bank of Cyprus or the Eurosystem.

Correspondence: Antonis A. Michis, Statistics Department, Central Bank of Cyprus, 80 Kennedy Avenue, P.O. Box 25529, CY-1395, Nicosia, Cyprus. Email: AntonisMichis@centralbank.gov.cy

1. INTRODUCTION

Ever since the seminal paper by Bates and Granger (1969), forecasting combinations have received considerable attention in the forecasting literature. Several combination methods have been proposed, ranging from simple averages of individual forecasts to sophisticated regression-based methods. Diebold (2004, p. 303) reviews some of the most important models in this field, emphasising that any regression tool is potentially applicable when the purpose is to combine forecasts. Most empirical studies tend to confirm that combining forecasts can improve forecasting accuracy (see, for example, the surveys by Clemen 1989 and Timmermann 2006), and usually simple averages provide satisfactory results (Makridakis and Winkler, 1983).

In parallel with the development of combination methods, several encompassing tests have been proposed in the literature (to be reviewed in section 2). These tests permit the evaluation of information inherent in individual forecasts by testing whether one forecast incorporates all of the relevant information inherent in rival forecasts. In an empirical study, Fang (2003) demonstrated that encompassing tests in the context of a forecasting combination exercise are usually complementary and must be evaluated jointly to reach a conclusion. However, with a large number of individual forecasts, a large number of encompassing tests must be performed, and as a result, inference becomes time consuming and complicated.

For this reason, several dimension reduction methods have been proposed in the literature that incorporate a selection step before combining forecasts to reduce the dimensionality of the forecasting combination problem. Dimension reduction procedures are useful because they eliminate forecasts that do not contribute towards reducing the forecast errors, and in this way, they improve the forecasting performance of the combination. Kisinbay (2010) proposed a useful algorithm in this direction that uses encompassing tests. However, despite the improvement in terms of time, the results are difficult to interpret. This complicates the algorithm’s usage in practice.

In this study, a semiparametric modelling framework is proposed that provides a useful method for evaluating forecasting combinations that is both computationally simple and efficient. By modelling the squared forecast errors as a function of the number of individual forecasts entering a combination, insights can be obtained regarding the number of individual forecasts to include in the combination and the marginal contribution of additional individual forecasts. The smooth partial regression estimates generated by semiparametric regression models provide two dimensional plots of the relationship between squared forecast errors and the number of forecasts entering the combinations.

These plots can facilitate decision making in situations that involve the integration of several different sources of forecasting information (see, for example, Sanders and Ritzman, 2004) and in situations where operations management requires the generation of forecasts for a large number of items (e.g., stock keeping units sold by an organisation). Franses (2011) describes one relevant case study of an organisation that generates forecasts for 1,211 products using forecasting combinations.

The rest of the article is organised as follows. Section 2 reviews the literature on forecast encompassing tests. Section 3 introduces the proposed semiparametric modelling method for monitoring forecasting combinations. Section 4 empirically evaluates the proposed method with six macroeconomic forecasts published by Her Majesty's Treasury (HM Treasury). Section 5 summarises the main conclusion of the study.

2. FORECASTING COMBINATIONS AND ENCOMPASSING TESTS

Chong and Hendry (1986) first proposed the use of encompassing tests to evaluate forecasting combinations by testing whether one model encompasses the forecasting properties of rival models. In its simplest version, an encompassing test can be carried out by running a regression of the realised values of a variable on forecasts generated by the individual models used in the forecasting combination. Simple t-tests of the regression coefficients can indicate whether one model encompasses the others, whether all models contain useful information about the future values of the variable, or whether neither model encompasses the others.

As suggested by Diebold (1989), encompassing can be seen more broadly as facilitating the combination of information sets, and the author recommended the exploration of combination methodologies that shrink the combining weights towards a measure of central tendency (e.g., the arithmetic mean). Fair and Shiller (1990) further elaborated the idea of combining information sets by proposing an encompassing test that provides better information than plain comparisons of relative mean squared errors (RMSE). Through an extensive empirical study, the authors found that combinations of structural and VAR macro-econometric models improve four-quarter-ahead forecasting of GDP changes.

The encompassing test proposed by Chong and Hendry has been extended in several other ways. Ericsson and Marquez (1993) proposed a generalised test that accounts for systematic biases in forecasts, the effects of multistep ahead forecasts, model nonlinearity, uncertainty in the coefficients, and facilitates simultaneous comparisons against several models. The results can also be used to improve model misspecification.

Another useful extension was provided by Harvey et. al. (1998), who proposed robust tests for the case of forecast error non-normality when comparing two forecasts. These were subsequently extended by Harvey and Newbold (2000) for comparisons of multiple forecasts made at horizons greater than one period, while accounting at the same time for forecast error heteroskedasticity and autocorrelation. Costantini and Kunst (2011) extended the work of Harvey and Newbold (2000) by demonstrating, through an extensive simulation study, how multiple encompassing tests can be used to determine the weights in forecast averaging.

Forecast encompassing tests can also prove useful in model specification, an aspect demonstrated in an empirical study by Fang (2003). Another empirical finding emphasised by the same author, and relevant to the subject of this article, is the conclusion that it is important to consider all relevant encompassing tests in the context of a specific project because the different tests may be complementary to one another.

When the number of individual forecasts entering a combination is large, evaluating and testing all possible combinations can prove time consuming and difficult. For this reason, a number of studies have proposed two-stage procedures for forecasting combinations. In the first stage, the dimensionality of the problem is reduced by selection of a subset of the available individual forecasts according to some criteria. The dimension reduction criteria usually involve statistical tests (e.g., encompassing tests) or model selection criteria (e.g., Bayesian information criteria). In the second stage, the selected forecasts are used to form the forecasting combination. Studies in this direction include Hallman and Kamstra (1989), Chen and Anandalingam (1990), Chandrasekharan et. al. (1994), Swanson and Zeng (2001) and Kisinbay (2010).

3. SEMIPARAMETRIC MODELLING OF FORECAST ERRORS

To investigate the relationship between the forecast errors and the number of individual forecasts used in a forecasting combination, a semiparametric additive regression model can be estimated as follows

$$E = a + \sum_{i=1}^{11} M_i + f(I) + \varepsilon. \tag{1}$$

The dependent variable (E) represents the squared forecast errors associated with the forecasting combination, irrespective of which combination method was used (e.g., a measure of central tendency or a regression-based method). The linear part of the model consists of eleven indicator variables (M), one for each month of the year except for January, which is captured by the intercept (a). The nonparametric part of the model (f) enters the model as an unknown smooth function of the number of individual forecasts used in the combination (I). The last term represents the error of the model (ε).

The monthly indicator variables are necessary in the context of the present study to reflect the impact of the time period of the forecast performance on the squared forecast errors and therefore the accuracy of the predictions. As will be explained in section 4, the data used for the empirical evaluation of model (1) consist of monthly forecasts of the end-of-year percentage changes of selected macroeconomic variables published by several organisations. Consequently, for each variable, there are 12 forecasting combinations available for each year. Furthermore, it is reasonable to expect that forecasts published towards the end of the year (e.g., in December), will be more accurate and have lower squared forecast errors than forecasts published earlier in the year (e.g., in

January). This is because more information is available to forecasting organisations towards the end of the year on which they can base their revisions of previous forecasts.

The semiparametric modelling framework proposed above can be generalised to include several other explanatory variables that likely influence the forecast errors. For example, Franses (2011) used a linear model to investigate the factors that influence expert forecasts and included lagged model-forecast errors and lagged values of the forecasted variable as explanatory variables. It is possible to incorporate similar variables in the linear part of model (1). However, because the data used in this study are not pure time series data, the specification with the indicator monthly variables was preferred.

Appendix A provides some technical details concerning the specification and estimation of semiparametric additive regression models that are estimated with the backfitting algorithm proposed by Hastie and Tibshirani (1990). This algorithm starts by estimating a linear regression model with the least squares method. This estimate is followed by smoothing of the partial residuals against each one of the independent variables to obtain new estimates of the partial regression functions. The procedure is then repeated every time there is a new partial regression estimate until convergence. In all of the applications presented in this study, smoothing was performed with local polynomial regression using kernel weights.

4. EVALUATION OF UK MACROECONOMIC FORECASTS

4.1. HM Treasury Data

The usefulness of the proposed method was evaluated using data for the period from May 2000 until December 2010, published on a monthly basis by HM Treasury (2000 – 2010) in the report “Forecasts for the UK economy: a comparison of independent forecasts”. This report provides a summary of published forecasts on key macroeconomic variables by several forecasting organisations. The selection of organisations included in the report is subject to review and changes in time.

Even though the report includes forecasts for many macroeconomic variables, not all forecasting organisations provided forecasts for all the variables. Furthermore, many organisations did not provide forecasts for all the months during the period May 2000 -

December 2010. After careful inspection of the available data (and sample sizes), six variables were selected for inclusion in the analysis: GDP, inflation (based on the retail price index), consumption, exports, imports and government spending. Following Kisinbay (2010), apart from sample size considerations, the choice of variables was also guided by the fact that these are perhaps the most well-known economic quantities to the general public that also receive the greatest media coverage in the UK.

The forecasts reported each month refer to end-of-year estimates of the percentage change in each variable. For example, the October 2010 report includes estimates published by the forecasting organisations regarding the percentage change of total GDP in 2010 in comparison to total GDP in 2009. In the case of inflation, the report includes forecasts of the percentage change from the final quarter in 2009 to the final quarter of 2010. Similarly, the November 2010 report will include the revised forecasts of the organisations regarding the end of year percentage changes in the variables. The revisions reflect the availability of new information while progressing towards the end of the year. As explained in section 3, this information effect on forecast accuracy is captured by the monthly indicator variables included in the linear part of model (1).

For each one of the selected variables, average forecasts were calculated for each month in the sample, using as inputs the individual forecasts of the organisations. This combination method was preferred because it is simple and the expected performance quality is supported by several empirical studies (see, for example, Makridakis and Winkler, 1983, and more recently, Jose and Winkler, 2008). However, because the number of organisations that submitted forecasts was not stable across time periods, changes were also observed in the number of individual forecasts used in the combinations.

Following the combination of individual forecasts, the squared forecast errors of the combinations (averages) for each month were calculated by obtaining the actual, end-of-year percentage changes in each variable from the publications of the UK Office for Statistics. These squared forecast errors were subsequently used as dependent variables in the semiparametric models developed according to equation (1). Table 1 includes summary statistics for the squared forecast errors and the number of individual forecasts used for calculating the forecast averages of each variable. Despite the fact that similar numbers of forecasts were used for each variable, the results for the squared forecast errors vary, particularly for exports and imports.

Table 1 Summary Statistics

Variable	Obs	Mean	Std Dev	Min	Max
GDP					
Squared forecast errors	128	0.625	1.245	0.001	8.381
Number of individual forecasts	128	41.266	2.041	37.000	44.000
Inflation					
Squared forecast errors	128	0.858	1.290	0.001	5.840
Number of individual forecasts	128	34.313	3.743	22.000	41.000
Consumption					
Squared forecast errors	128	5.983	4.821	0.002	21.623
Number of individual forecasts	128	41.195	2.012	37.000	44.000
Exports					
Squared forecast errors	128	70.913	76.325	0.001	472.059
Number of individual forecasts	128	39.242	2.121	34.000	43.000
Imports					
Squared forecast errors	128	91.176	92.708	0.001	443.030
Number of individual forecasts	128	39.242	2.110	34.000	43.000
Goverment Spending					
Squared forecast errors	128	6.525	16.198	0.001	63.640
Number of individual forecasts	128	40.789	1.947	36.000	44.000

4.2. Results

Estimation of the semiparametric models was performed with the backfitting algorithm, which produced estimates of the coefficients (linear part) and of the smooth partial regression functions (nonparametric part) in each model. The coefficient estimates for the monthly indicator variables are included in Table 2 with their standard errors.

Table 2 Estimation of Linear Components

Variable	GDP	INF	CON	IMP	EXP	GSP
Intercept	1.483*	1.370*	6.796*	120.948*	107.446*	5.976
	(0.384)	(0.363)	(1.322)	(30.079)	(24.223)	(4.915)
February	-0.471	0.174	-0.117	-2.365	-10.038	0.895
	(0.542)	(0.514)	(1.870)	(42.539)	(34.256)	(6.951)
March	-0.654	-0.001	-0.572	-12.810	-21.013	1.457
	(0.542)	(0.514)	(1.870)	(42.539)	(34.256)	(6.951)
April	-0.753	-0.256	-1.028	-19.809	-31.697	0.995
	(0.542)	(0.514)	(1.870)	(42.539)	(34.256)	(6.951)
May	-0.879**	-0.444	-0.272	-30.090	-36.804	2.297
	(0.530)	(0.502)	(1.827)	(41.561)	(33.469)	(6.791)
June	-0.874**	-0.588	-0.184	-31.513	-37.185	1.886
	(0.530)	(0.502)	(1.827)	(41.561)	(33.469)	(6.791)
July	-0.930**	-0.582	-0.137	-34.272	-39.078	1.489
	(0.530)	(0.502)	(1.827)	(41.561)	(33.469)	(6.791)
August	-1.003**	-0.515	-0.268	-35.035	-43.506	1.552
	(0.530)	(0.502)	(1.827)	(41.561)	(33.469)	(6.791)
September	-1.033*	-0.641	-1.021	-41.137	-49.358	0.971
	(0.530)	(0.502)	(1.827)	(41.561)	(33.469)	(6.791)
October	-1.168*	-0.848**	-1.402	-44.998	-57.323	-1.493
	(0.530)	(0.502)	(1.827)	(41.561)	(33.469)	(6.791)
November	-1.185*	-1.018	-1.860	-46.570	-55.489	-1.317
	(0.530)	(0.502)	(1.827)	(41.561)	(33.469)	(6.791)
December	-1.212*	-1.238	-2.755	-51.027	-49.329	-2.049
	(0.530)	(0.502)	(1.827)	(41.561)	(33.469)	(6.791)

Significant: * 5%, **10% ; Standard errors are in parentheses.

Each column in the table corresponds to one of the model variables: gross domestic product (GDP), inflation (INF), consumption (CON), imports (IMP), exports (EXP) and government spending (GSP). It can be observed that as one moves to months that are closer to the end of the year, the coefficients have larger negative values, indicating bigger reductions in the squared errors of the combined forecasts. These reductions reflect the availability of more information regarding the state of the economy towards the end of the year.

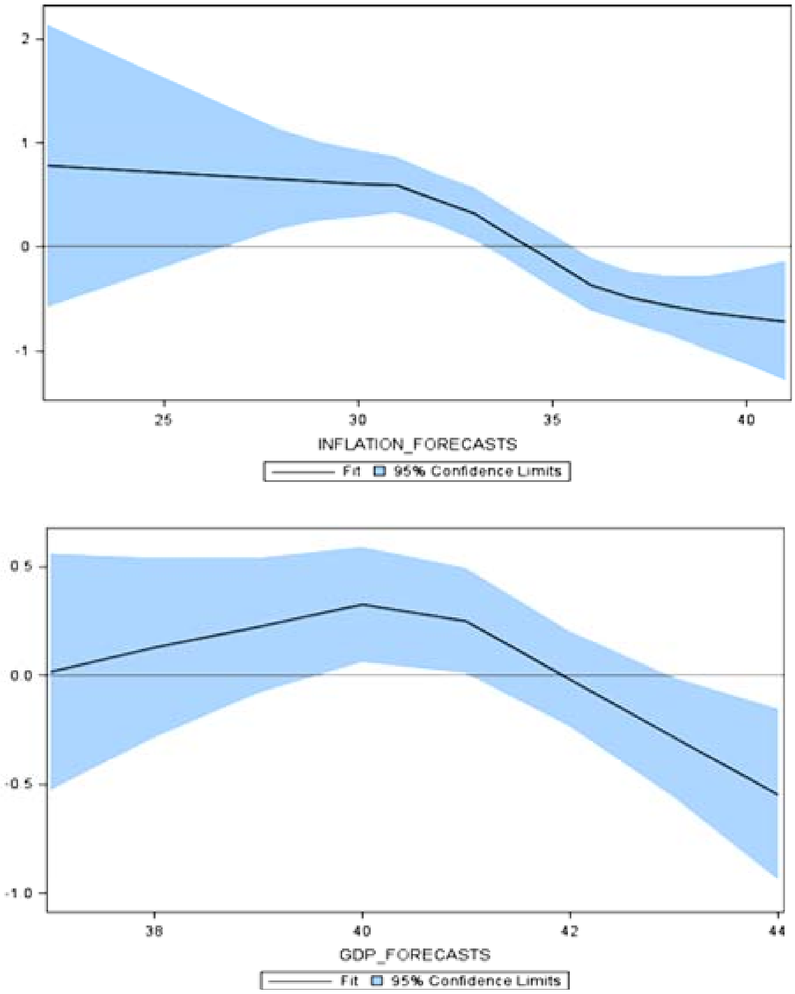
The smooth partial regression functions in the models were estimated by incorporating local polynomial regression smoothing in the backfitting algorithm. Information regarding these estimates is included in Table 3. The smoothing parameter (or bandwidth) controls the smoothness of the partial regression functions and is adjusted to include a fixed proportion of the data around a focal point (see the appendix for more details). The associated degrees of freedom (DF) and F-test for each variable are used to test the contribution of the number of individual forecasts (the explanatory variable) in explaining the dependent variable in the semiparametric model. In all cases, the explanatory variables provide a statistically significant addition to the model at the 5% significant level except in the case of imports where the contribution is significant at the 18% level.

Table 3 Estimation of Smoothing Components

Model	Smoothing Parameter	DF	F value	Pr > F
GDP	0.988	1.458	6.810	0.004
INF	0.738	2.009	16.270	0.001
CON	0.809	2.030	26.370	0.002
EXP	0.801	1.965	3.600	0.031
IMP	0.809	1.983	1.740	0.181
GSP	0.816	1.950	12.220	0.001

The two dimensional plots of the smooth partial regression functions, generated by model (1), can be used to assess how changes in the number of individual forecasts used in the forecasting combinations influence the squared forecast errors. Figure 1 demonstrates the estimated smooth functions for inflation and GDP. For both variables there are significant reductions in the squared forecast errors (measured on the vertical axis) as the number of individual forecasts used in the calculation of the forecast averages increases (measured on the horizontal axis). This is more pronounced with the addition of more than 30 forecasts in the case of inflation and more than 41 forecasts in the case of GDP.

Figure 1 Partial Regression Functions: Inflation and GDP



The shaded areas around the smooth functions represent 95% confidence intervals, which were calculated based on the procedure proposed by Cleveland, Devlin and Grosse (1988) for local polynomial regression smoothing. The intervals are wider for large and small values of the explanatory variables, reflecting the availability of small sample sizes in the neighbourhood of these values in the data. Another useful insight included in these plots concerns the reduction in squared forecast errors when the number of available forecasts increases above specific thresholds. This information can be used to evaluate whether the inclusion of previously unavailable individual forecasts improves the combination accuracy by introducing new information. It would be time consuming and less straightforward to assess the information content of new individual forecasts using all possible combinations of encompassing tests.

Figure 2 Partial Regression Functions:
Consumption, Government Spending, Exports and Imports

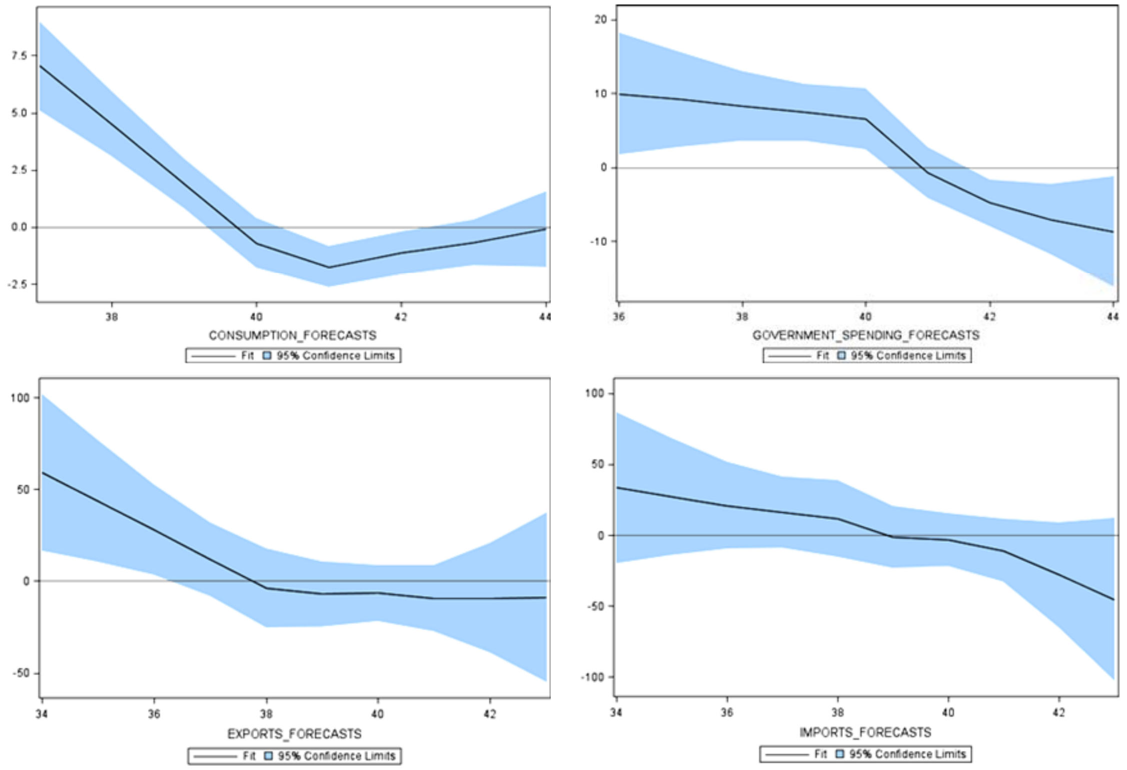


Figure 2 includes two dimensional plots of the partial regression functions for the other variables. For government spending and imports, significant reductions in the squared forecast errors are achieved with the inclusion of more than 40 and 38 individual forecasts in the forecasting combinations, respectively. This is not true for consumption forecasts where the addition of more than 41 forecasts did not help in reducing the squared forecast errors. This indicates that the additional forecasts did not provide any new information in the forecasting combinations but, on the contrary, have decreased the combination forecast accuracy. In this case, it is recommended to include in the forecasting combination only the 41 forecasts that generated the smallest squared forecast errors. For exports, there are only marginal reductions in the squared forecast errors when the combination includes more than 38 individual forecasts. In this case, it is recommended to include all of the forecasts in the calculation of the combination and to examine other, potentially beneficial methods for calculating the combining weights.

5. CONCLUSIONS

Forecasting combinations are used by many forecasting organisations because both theoretical and empirical evidence suggest that they can provide substantial improvements in forecast accuracy. At the same time, it is important to evaluate the information content in these combinations with proper statistical criteria to choose the best possible sets of individual forecasts for inclusion in the combinations. When the number of available individual forecasts is large, this evaluation can be time consuming and difficult because most existing encompassing tests require many calculations and considerable effort to evaluate all of the results.

The semiparametric modelling procedure proposed in this paper provides a computationally simple and fast method for evaluating forecasting combinations using two dimensional plots of the relationship between squared forecast errors and the number of individual forecasts used in the forecasting combinations. The proposed method will be useful to organisations that routinely perform forecasts for large numbers of items and utilise combinations of many different sources of information.

REFERENCES

Bates, J. M. and Granger, C.W.J. (1969). “The combination of forecasts”. Operations Research Quarterly, 20: 451–468.

Chandrasekharan, R., Moriarty, M.M. and Wright, G.P. (1994). “Testing for unreliable estimators and insignificant forecasts in combined forecasts”. Journal of Forecasting, 13: 611–624.

Chen, L. and Anandalingam, G. (1990). “Optimal selection of forecasts”. Journal of Forecasting, 9: 283–297.

Chong, Y.Y. and Hendry, D.F. (1986). “Econometric evaluation of linear macro-economic models”. Review of Economic Studies, 53: 671-690.

Clemen, R. T. (1989). “Combining forecasts: A review and annotated bibliography”. International Journal of Forecasting, 5:559–583.

Cleveland, W.S., Devlin, S.J. and Grosse, E. (1988). “Regression by local fitting”. Journal of Econometrics, 37: 87-114.

Costantini, M. and Kunst, R.M. (2011). “Combining forecasts based on multiple encompassing tests in a macroeconomic core system”. Journal of Forecasting, 30: 579-596.

Diebold, F.X. (1989). “Forecast combination and encompassing: reconciling two divergent literatures”. International Journal of Forecasting, 5: 589-592.

Diebold, F.X. (2004). Elements of Forecasting. 3rd ed. Cincinatti, OH: South-Western.

Ericsson, N.R. and Marquez, J. (1993). “Encompassing the forecasts of U.S. trade balance models”. The Review of Economics and Statistics, 75: 19-31.

Fair, R.C. and Shiller, R.J. (1990). “Comparing information in forecasts from econometric models”. American Economic Review, 80: 375-389.

Fang, Y. (2003). “Forecasting combination and encompassing tests”. International Journal of Forecasting, 19: 87-94.

Fox, J. (2000). Multiple and Generalized Nonparamatric Regression. Thousand Oaks, CA: Sage.

Fox, J. (2008). Applied Regression Analysis and Generalized Linear Models. Thousand Oaks, CA: Sage.

Franses, P.H. (2011). “Averaging Model Forecasts and Expert Forecasts: Why Does It Work?”. Interfaces, 41: 177-181.

Harvey, D. and Newbold, P. (2000). “Tests for multiple forecast encompassing”. Journal of Applied Econometrics. 15: 471-482.

Harvey D.I., Leybourne S.J., Newbold P. (1998). “Tests for forecast encompassing”. Journal of Business and Economic Statistics 16: 254-259.

Hastie, T.J. and Tibshirani, R.J. (1990). Generalized Additive Models. London: Chapman & Hall.

Hallman, J. and Kamstra, M. (1989). “Combining algorithms based on robust estimation techniques and co-integrating restrictions”. Journal of Forecasting, 8: 189–198.

HM Treasury (2000 - 2010). Forecasts for the UK economy: a comparison of independent forecasts. Available on the Treasury’s website: http://www.hm-treasury.gov.uk/forecasts.

Jose, V.R.R. and Winkler, R. (2008). “Simple robust averages of forecasts: Some empirical results”, International Journal of Forecasting, 24: 163-169.

Kisinbay, T. (2010). “The use of encompassing tests for forecast combinations”. Journal of Forecasting, 29: 715-727.

Makridakis, S. and Winkler, R. (1983). “Averages of forecasts: some empirical results”. Management Science, Vol. 29, pp. 987-96.

Sanders, N.R. and Ritzman, L.P. (2004). “Integrating judgmental and quantitative forecasts: Methodologies for pooling marketing and operations information”. International Journal of Operations and Production Management, 24: 514–529.

Swanson, N.R. and Zeng, T. (2001). “Choosing among competing econometric forecasts: regression-based forecast combination using model selection”. Journal of Forecasting, 20: 425–440.

Timmermann, A. (2006). “Forecast combinations”. In Elliott, G., Granger, C. W. J. and Timmermann, A. eds. Handbook of Economic Forecasting, Vol. 1. Elsevier, Amsterdam, 135–196.

APPENDIX: SEMIPARAMETRIC ADDITIVE REGRESSION

In additive regression models, the conditional mean value of the dependent variable is specified as the sum of univariate, smooth functions of several explanatory variables

$$E(y \mid x_1, x_2, ..., x_k) = a + f_1(x_{i1}) + f_2(x_{i2}) + ... + f_k(x_{ik}).$$

The advantage of this specification is that it reduces to a series of two-dimensional nonparametric, partial-regression models (Fox 2000, p.27). These can be estimated with an appropriate scatterplot smoother and each partial regression can be represented in a two dimensional plot, while holding the other explanatory variables constant. This greatly facilitates the visual inspection and interpretation of the relationships between the dependent variable and the explanatory variables. Semiparametric regression models can be considered as special cases of the additive regression model where some of the explanatory variables enter linearly and can include indicator variables

$$E(y \mid x_1, x_2, ..., x_k) = a + \beta_1 x_{i1} + ... + \beta_r x_{ir} + f_{r+1}(x_{ir+1}) + ... + f_k(x_{ik}).$$

Additive regression models can be fit to data with the backfitting algorithm proposed by Hastie and Tibshirani (1990). This algorithm starts with preliminary estimates of the partial regression functions based on the following linear regression model, which is estimated using the least squares method

$$y_i - \bar{y} = b_1(x_{i1} - \bar{x}_1) + b_2(x_{i2} - \bar{x}_2) + ... + b_k(x_{ik} - \bar{x}_k),$$

and where $\hat{f}_r^0(x_{ir}) = b_r(x_{ir} - \bar{x}_r)$ is the initial partial regression estimate for the independent variable x_r. In a second step, the algorithm forms the partial residual

$$e_{ir} = y_i - \bar{y} - \sum_{j \neq r} b_j(x_{ij} - \bar{x}_j)$$

and proceeds to smooth it against x_{ir} to obtain a new estimate of the partial regression function $\hat{f}_r^1$. This estimate is then used to calculate the partial residuals and smooth functions of the other explanatory variables. The procedure is repeated again for all variables by iteratively smoothing the partial residuals for each explanatory variable, each time using the latest available estimates of the smooth functions for the other variables. The algorithm continues repeating the procedure until convergence is achieved and the partial regression functions stabilise (Fox, 2000, p. 32).

Scatterplot smoothing to estimate the partial regression functions is usually performed using local polynomial regression or smoothing splines. In all the applications in this study, the local polynomial regression method was used, which fits the following equation to a sample of observations around the focal point x_0 of the x_r variable,

$$y_i = A + B_1(x_{ir} - x_0) + B_2(x_{ir} - x_0)^2 + ... + B_p(x_{ir} - x_0)^p$$

by minimising the weighted residual sum of squares $\sum_{i=1}^{n} w_i E_i^2$ (weighted least squares estimation with a sample of n observations). The fitted value at the focal x_0 is then $\hat{y} \mid x_0 = A$. Estimation is repeated at several other representative focal values to derive the smooth function of the explanatory variables x_r. The weights (w_i) used in the estimation procedure are based on the kernel function that gives greater weights to observations near the focal x_0 and declines as the distance from the focal point increases

$$w_i = K[(x_i - x_0)/h].$$

Of particular importance is the bandwidth (or smoothing) parameter h, which controls the smoothness of the partial regression function. Larger values of the parameter provide smoother estimates. The bandwidth is usually adjusted to include a fixed proportion of the data around the focal point (a method called nearest neighbour bandwidth), and this adjustment was used for the models in this study. The specific value of the bandwidth parameter for each model was selected following the suggestions in Fox (2008, p.482), according to which an initial estimate can be obtained automatically with cross validation. Cross validation is then followed by visually guided trial and error to achieve the smallest value of the parameter that provides a reasonably smooth fit to the data.

Statistical inference and procedures for estimating confidence intervals for the partial regression functions are described in Fox (2008, p. 513-514). The confidence intervals presented in the figures of this study were calculated based on the procedure proposed by Cleveland, Devlin and Grosse (1988) for local polynomial regression smoothing. The contribution of each explanatory variable in an additive regression model can be tested with an F-test by comparing the full additive model with a model omitting the explanatory variable in question

$$F = \frac{\frac{RSS_0 - RSS_1}{df_1 - df_0}}{\frac{RSS_1}{df_{res}}}.$$

RSS_1 represents the residual sums of squares for the full model that has df_1 degrees of freedom. RSS_0 is the residual sum of squares for the model omitting the explanatory variable, and this model has df_0 degrees of freedom. Finally, $df_{res} = n - df_1$ represents the error degrees of freedom.