



ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

International Journal of Forecasting xx (2005) xxx–xxx

www.elsevier.com/locate/ijforecast

Discussion

Comments on: “Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A re-examination”

Alfonso Novales

Departamento de Economía Cuantitativa, Universidad Complutense-Madrid, Spain

1. Introduction

The paper presented by Professors Terasvirta, van Dijk, and Medeiros (TvDM) is a very thorough and complete discussion on forecast evaluation of smooth transition autoregression (STAR) and neural network (NN) models using monthly macroeconomic time series data. As mentioned in the introduction to the paper, a fair amount of work has already been done suggesting a nonsignificant gain from using NN models for forecasting, while much less work has been done evaluating the forecasting performance of STAR models. The paper contributes by using an international macroeconomic data set, which has not been examined from the perspective of nonlinear forecasting. The structure of the data set is interesting, allowing for possible regularity patterns to emerge across variables and countries. Such regularities would be very useful as a guide to detect cases in which nonlinear forecasting models should be considered.

STAR and NN models are both very flexible parameterizations that should be able to capture many types of nonlinearities in the data. They are not with-

out problems, since they need a very careful specification regarding the transition rule in the case of STAR models and the number of hidden units in the NN model. Except for their simplest versions, both models contain a large number of parameters, leading to flat likelihood surfaces and a consequent poor identification, as an indication of overparameterization. This could be considered the curse of nonlinearity: a simple nonlinear specification may be too close to linearity, but a more interesting model may be hard to identify and estimate with precision.

The paper makes an interesting reading for forecasting practitioners, because the authors discuss a number of very relevant issues on forecasting with nonlinear models and describe the main references on each topic: (i) Do linearity tests provide a reliable guide to post-sample forecast accuracy? Should they guide model specification? (ii) Do we need to use different models for different forecast horizons as it is usually done? (iii) How should forecast performance be evaluated? Is the RMSE appropriate for nonlinear forecasting models? (iv) In the presence of model uncertainty, how useful are forecast combinations? Other issues, like: (v) What is the more appropriate data transformation for forecasting? (vi) How should we deal with seasonality? (vii) Should we

E-mail address: anovales@ccee.ucm.es.

58 correct for outliers prior to forecasting?, which were
59 mentioned in previous versions of the paper have been
60 reasonably left aside for future research, while in the
61 TvDM paper specific choices have been made for
62 each of them: all variables are handled in annual
63 growth rates, except for interest rate and the rate of
64 unemployment, seasonal dummies are employed, and
65 the time series were corrected for outliers.

66 2. Comparing forecasting performance

67 Producing an informal ranking of competing fore-
68 cast specifications using observed RMSE values may
69 lead to choosing a preferred model which is essentially
70 equivalent to alternative specifications with slightly
71 higher RMSE values. It is therefore convenient to
72 run formal tests for statistical significance of RMSE
73 differences. For sound reasons, the Diebold-Mariano
74 (DM) approach has quickly become a standard proce-
75 dure, but using it as the single criterion for model
76 choice may still be risky, because of the imperfect
77 relationship between quantitative relevance and statisti-
78 cal significance. It may be possible to find statisti-
79 cally different RMSE values for forecast paths, which
80 could provide the analyst with a very similar qualita-
81 tive impression on the future evolution of the variable
82 under study. On the other hand, a low precision in
83 estimation might lead to concluding that RMSE values
84 emerging from two forecast paths with dissimilar char-
85 acteristics are not statistically different from each
86 other. In general, the standard approach of summariz-
87 ing all the information from a predicted trajectory on a
88 single number may not be the best way to proceed.

89 Let us examine the forecasting results in TvDM.

90 2.1. The more informal approach

91 Even under a purely numerical comparison, it is
92 questionable that differences of the order of 1% in
93 RMSE values should be taken into account. The
94 extensive results in Table 2 could be summarized
95 by considering a 5% (alternatively, a 2.5%) differ-
96 ence in RMSEs as an approximate threshold for
97 numerical relevance.¹ Out of 28 pair comparisons (4

forecast horizons and 7 countries), the number of 98
cases in which the RMSE for the STAR model is at 99
least 5% (or, alternatively, 2.5%) less than the one for 100
the linear AR model appears in the first row of the 101
table²: 102

Number of cases with an RMSE 5% (2.5%) below alternative model	IP	CPI	M1	STIR	VEX	VIMP	UR	Total	t.1
STAR	9 (13)	1 (2)	13 (17)	4 (6)	0 (2)	3 (4)	4 (6)	34 (50)	t.1.2
Linear AR	0 (0)	3 (5)	1 (2)	12 (15)	5 (9)	2 (5)	4 (4)	27 (40)	t.1.3

But the comparison might be misleading unless it 103
is performed in both directions, so the second row in 104
the table shows the number of cases in which the 105
RMSE for the linear AR model is at least 5% lower 106
than the one for the STAR model.³ This is important: 107
applying the 5% criterion to consumer prices, 108
imports or the unemployment rate, the STAR 109
model does not beat the linear AR model in fore- 110
casting, but neither does the AR model beat the 111
STAR model. It seems impossible to say that either 112
model is best, the comparison being clearly variable- 113
specific. For industrial production and the money 114
supply, the STAR model produces better forecasts, 115
while for interest rates and possibly exports the 116
linear AR model might be preferred to the STAR 117
model. The same evaluation would be reached using 118
a 2.5% reduction in RMSE as a threshold for sig- 119
nificance,⁴ as shown by the figures in brackets in the 120
table. If we use 1% differences between RMSE 121
values, a very similar impression still arises, although 122
the risk is now high that we compute as significant a 123
difference between RMSEs, which may be purely due 124
to sampling error. This seems to be the case with 125
consumer prices, which would give the impression 126
of being better predicted by the linear AR model 127
under this criterion. 128

² The number of comparisons for the unemployment rate is 20, rather than 28.

³ Which amounts to an entry either above 1.05 for STAR model in Table 2.

⁴ Using entries either below 0.975 or above 1.025 for STAR model in Table 2.

¹ This will always be a tough choice without examining the whole predicted path.

132 The comparison between the NN and the linear AR
133 models using 2.5% RMSE differences generates,

t2.1	Number of cases with an RMSE 2.5% below alternative model	IP	CPI	M1	STIR	VEX	VIMP	UR	Total
t2.2	NN	11	7	0	3	2	10	6	39
t2.3	Linear AR	0	19	14	19	18	3	1	74

134 showing the better edge of the NN model for indus-
135 trial production, imports, and the unemployment rate,
136 while the linear AR model dominates for all other
137 variables. A similar analysis to compare the perfor-
138 mance of STAR and NN models produces:

t3.1	Number of cases with an RMSE 2.5% below alternative model	IP	CPI	M1	STIR	VEX	VIMP	UR	Total
t3.2	STAR	10	19	22	16	12	7	7	93
t3.3	NN	5	7	0	5	1	9	11	38

139 The STAR model seems to perform better than the
140 NN model for all variables except imports and the rate
141 of unemployment, for which both models seem to
142 produce forecast of similar quality. The result is never-
143 theless of small relevance for interest rates and
144 exports, since both models are then beaten by the
145 linear AR.

146 Overall, the linear AR model seems to be better
147 for interest rates and exports, as mentioned, the
148 STAR model should be preferred for industrial pro-
149 duction and money, while the NN model seems to
150 forecast better than the two alternatives for imports
151 and the unemployment rate. Either the linear AR or
152 the STAR models should be used for consumer
153 prices. It is hard to think of a more equilibrate result,
154 but the main issue is whether this ranking of fore-
155 casting performance across variables is robust over
156 time and across a wider sample of countries, which
157 would need of further analysis. Even more important,
158 the main question for further research is to figure out
159 the specific statistical characteristics in a given vari-
160 able that produces the described ranking among
161 forecasting specifications.

2.2. Statistical significance of forecast differences 162

The previous analysis uses an informal count of
RMSE differences above a given threshold, with no
formal statistical test for significance of RMSE differ-
ences. The Diebold-Mariano (DM) approach is fol-
lowed in the paper to provide statistical significance
thresholds for RMSE differences, but it is somewhat
disturbing that the results obtained applying the DM-
test or comparing RMSE values may be so different.

Our discussion above on the relevance of RMSE
comparisons suggests two possible conflicting results:
one, that RMSE ratios close to 1.0, which suggest
very similar predicted paths, may come together with
rejections in the DM-test. Second, that a RMSE ratio
well below 1.0 may come together with no rejection
of the null hypothesis of equal forecasts in the DM-
test. As an example of the first situation, RMSE ratios
of 1.009, 1.019, 1.124, 1.000, and 0.968 for the
STAR/AR comparison when predicting the unemploy-
ment rate at the 1-month horizon lead to two
cases in which the linear model is preferred, for
zero cases in which the STAR model is preferred.
Similarly, RMSE ratios of 1.000, 1.061, 1.000, 0.986,
1.018, 1.026, and 1.000 for the STAR/AR compar-
ison when predicting CPI at the 12-month horizon
lead to two cases in which the linear AR model is
preferred.

Alternatively, as an example of the second situa-
tion, RMSE ratios of 0.871, 0.971, 0.783, 1.000,
1.000, 1.000, and 0.902 for the STAR/AR compar-
ison when predicting industrial production at the 12-month
horizon lead in the DM-test to only one case in which
the STAR model is preferred to the linear AR model.
There is no much to object to classical hypothesis
testing, but one would expect that reductions above
10% in RMSE should amount to a significant differ-
ence in the forecast path.

With the quadratic loss function standard in the
literature, the DM statistic is essentially the difference
of RMSE values for the two models being compared,
divided by the sample standard deviation of forecast
differences. A small variance for forecast differences
can explain that the DM-tests reject the null of equal
forecasts, while a large variance can explain that a
relatively large RMSE difference does not lead to
rejection in the DM-tests. It would be interesting to
figure out the reasons for forecast differences to be

209 measured with more or less precision across variables
 210 and models, but it is unclear that a possible lack of
 211 precision in estimating single period forecast differ-
 212 ences should be so crucial in establishing significant
 213 differences between forecast paths. While a test like
 214 DM introduces statistical rigor into RMSE compari-
 215 sons, what the forecasting analyst cares about is
 216 whether two alternative models lead to numerically
 217 different forecast paths, with distinct qualitative impli-
 218 cations. A minor numerical difference between fore-
 219 casts is not relevant for the analyst, even if it turns out
 220 to be statistically significant.

221 *2.3. Bayesian regularization*

222 One of the points discussed in the TvDM paper
 223 focuses on the relevance of Bayesian regularization
 224 incorporated into the NN model to achieve a neural
 225 network NN parameterization. The comparison
 226 between these two models along the previous lines
 227 yields,

t4.1	Number of cases with an RMSE 2.5% below alternative model	IP	CPI	MI	STIR	VEX	VIMP	UR	Total
t4.2	NN	15	17	16	8	7	12	12	87
t4.3	AR-NN	0	6	1	15	10	3	1	37

228 suggesting that Bayesian regularization⁵ is helpful
 229 for predicting industrial production, prices, money,
 230 imports, and the unemployment rate, but it per-
 231 forms poorly when dealing with interest rates and
 232 exports. The comparison is again variable specific,
 233 although if a global choice had to be done, the
 234 Bayesian regularization alternative seems preferable,
 235 the AR-NN being then last in a hypothetical order-
 236 ing of the forecasting models considered in the
 237 paper.

238 The picture emerging from this analysis is consis-
 239 tent with that from the DM approach followed in
 240 TvDM, except for a more clear dominance of STAR
 241 over neural network models as well as that of NN over
 242 the AR-NN model than it is stated in Section 7.1.

⁵ Pruning a large network, rather than adding layers to an initially small network.

3. Forecasting horizon and performance 243

It does not seem to be very interesting to compare 244
 forecasting performance for a given country, since 245
 whatever nonlinearity there is in the data set, it should 246
 be more a property of some variables than a country- 247
 specific characteristic. It seems quite more interesting 248
 to make the comparison for different forecast hori- 249
 zons. In that respect, the comparison between the 250
 STAR and linear AR models leads to, 251

	1	3	6	12	Total	t5.1
Number of cases with an RMSE 2.5% below alternative model						
STAR	0 (6)	9 (14)	12 (14)	13 (17)	34 (51)	t5.2
Linear AR	6 (9)	2 (8)	9 (10)	10 (13)	27 (40)	t5.3

Being a two-way comparison, there is no way to 252
 know whether the better relative forecasting behaviour 253
 of the STAR model relative to the linear AR model in 254
 the longer horizons is due to a gradual improvement in 255
 the former, a deterioration in the latter or both. But the 256
 table suggests that any possible gain from the non- 257
 linearity in the STAR model might come up for 258
 medium-term forecasts, with very few gains in 259
 short-term forecasting relatively. Analyzing the 260
 robustness of this relationship between relative per- 261
 formance and horizon and its possible causes seems to 262
 be an interesting issue for further research. 263

4. Pre-testing for linearity 264

Precisely because forecasting performance is vari- 265
 able-specific, with few dominance results across mod- 266
 els, it would be most helpful to have available 267
 statistical tools that may lead the analyst to decide 268
 between linear or nonlinear forecasting specifications. 269
 It is a good idea that linearity tests are used as such a 270
 tool, but the implied results do not seem to lead to the 271
 optimistic statements that can be read at some points 272
 in the paper. 273

For variables like industrial production, there 274
 seems to be a satisfactory consistency between the 275
 results of linearity tests and the relative forecasting 276
 performance of linear and nonlinear models. How- 277
 ever, for some other important variables like interest 278
 rates or prices, the opposite seems to be true. Except 279

280 in the US, neither the STAR nor the AR-NN speci-
 281 fications seem to capture the nonlinearity in interest
 282 rates that appears in Table 1 with both specifications
 283 as alternative hypothesis. The lower panel in that table
 284 shows clear rejection of linearity against an AR-NN
 285 alternative, but, again, RMSE values in Table 2 are
 286 nowhere suggestive of a clear preference for the for-
 287 mer model versus the linear AR model for forecasting
 288 purposes. Further analysis and design of better statis-
 289 tical procedures to lead in the choice of forecasting
 290 model is clearly needed.

291 5. Forecast combinations

292 Table 7 shows how there are some instances in
 293 which a combination of forecasts produces a smaller
 294 RMSE than the reference model, although reduc-
 295 tions in RMSE values are often small. Using again
 296 a 2.5% reduction in RMSE as an informal signifi-
 297 cance threshold, the NN+STAR combination seems
 298 to be best, lowering RMSE relative to the baseline
 299 model in almost 33% of the cases (variable, forecast
 300 horizon, country). All others show smaller improve-
 301 ments, the AR+NN and AR-NN+STAR combina-
 302 tions improving forecasts relative to the linear model
 303 in about 20% of the comparisons, the AR+STAR com-
 304 bination in 15% of the cases and the AR+AR-NN in
 305 10% comparisons.

306 Gains are more often obtained when predicting
 307 industrial production or the unemployment rate, and
 308 they also arise for prices and money. They are
 309 obtained less often for interest rates, exports, and
 310 imports. However, in most cases, forecast gains
 311 from using the NN+STAR combination are not as
 312 large as by using STAR model by itself, so that,
 313 although the combination beats the univariate linear
 314 autoregression model, it does not perform better than
 315 the best model in the combination. At least with this
 316 data set, it seems more promising to try to solve the
 317 hard problem of selecting the most appropriate fore-
 318 casting model for a given variable, rather than relying
 319 on the improvement that might arise by combining
 320 forecasts from different models.

321 Another issue is whether the encompassing tests
 322 provide some useful guidance in choosing good fore-
 323 cast combinations. The AR-NN+AR and AR-
 324 NN+STAR combinations have the higher number of

rejection cases for the encompassing tests in Table 4,
 but, as described above, they do not beat the baseline
 linear model very often. On the other hand, the pre-
 ferred NN+STAR combination does not come as pro-
 ducing a striking number of rejections in Table 4.
 Hence, the connection between forecasting encom-
 passing tests and the improvement produced by com-
 bining forecasts is not very tight, and this deserves to
 be considered in further research. The forecasting
 analyst faces ex-ante choices and cannot afford to
 wait for the actual data to be produced in order to
 do an ex-post evaluation of the best forecast combina-
 tion. As it is the case when choosing a forecasting
 model, what is crucial to the analyst is to have statis-
 tical procedures that might lead to a good choice of
 forecast combination.

6. Final considerations

Much effort is being done to impose nonlinear
 models on a given sample, often without having char-
 acterized previously in detail which aspects of the data
 we want to capture with the nonlinear specification.
 The real challenge is whether there is any chance that
 we could anticipate which variables are more likely to
 be better forecasted from a nonlinear than from a
 linear model and, as shown by TvDM, that is hard
 to emerge from general linearity tests.

We may need to change the approach. We should
 start with a detailed description of the main statistical
 characteristics of the data, which might lead into the
 search for a plausible nonlinear specification. The fact
 that forecast improvement from nonlinear models
 concentrates on a few variables suggests that there
 are specific characteristics producing that forecast
 gain. RMSE comparisons in the paper suggest that,
 over the set of variables and countries considered, the
 STAR model is a better nonlinear forecasting specifi-
 cation for some variables, while the NN model is
 better for imports and the unemployment rate. Hard
 as it may be, it would be most useful to further explore
 which characteristics of these two variables are behind
 this result.

Relative to the specific analysis in TvDM, it might
 be the case that macroeconomic data do not make a
 good case for nonlinearity. Is the type of data for
 which we would expect to see the type of features

370 associated to STAR and NN models? For instance,
371 STAR models, which produce the best results among
372 nonlinear specifications in the TvDM paper, can give
373 raise to a wide variety of structures: they can produce
374 smooth time varying parameters, but also drastic
375 changes in regime, which occur occasionally. The
376 question is whether we often have time variation in
377 linear representations of macroeconomic variables,
378 and of what type. Does it come as a gradual change
379 in parameter, or as a change of regime jump?

380 The example most often used in Econometrics
381 when referring to possible nonlinear stochastic struc-
382 tures it is that of frequently observed financial vari-
383 ables, which have features very different from those of
384 macroeconomic data. There are many examples in the
385 analysis of portfolio choice in different markets, in
386 which a few factors are supposed to explain a good
387 deal of the variation in a large vector of returns. To
388 estimate these relationships with frequently sampled
389 data, as well as to forecast high frequency returns, the
390 nonlinear specifications considered in the TvDM
391 paper might be quite useful.

392 Another consideration deals with summarizing
393 forecasting performance. There is clearly a need to
394 depart from model searching using a single statistic
395 that summarizes forecasting performance over a given
396 horizon. The Diebold-Mariano approach takes this
397 into account, but we might want to go even further.
398 By using the full forecast path in detail, we may be
399 able to see whether a possible single large error is
400 contaminating the comparison, whether forecast errors
401 are correlated over time, and so on. As mentioned
402 above, in spite of the unquestionable statistical interest
403 of the DM approach, it is debatable whether the fore-
438

casting practitioner should care about statistical sig- 404
nificance or purely numerical differences in forecast 405
paths from alternative models. 406

407 But even before comparing forecasts, we should
408 evaluate the differences between estimated models.
409 For instance, it would be interesting to compare the
410 type of parameter time variation implied by estimated
411 alternative nonlinear models. More generally, we
412 could think of tests comparing the full set of residuals
413 from linear and nonlinear alternatives, or between
414 alternative nonlinear specifications, to see the extent
415 to which such models are different. For instance, it
416 might happen that the difficulties faced when specify-
417 ing the number of hidden layers in neural networks or
418 the number of lags in the shifting function in STAR
419 models could lead to a possibly overparameterized
420 model. If that is the case, residuals from both the
421 general and the restricted model might well be very
422 similar, and forecasts will likely be quite similar.

423 Regarding forecast combination, I am less optimistic
424 than the authors, for reasons explained above.
425 Encompassing tests are not very tightly related to
426 possible gains from forecast combination, the same
427 way that linearity tests do not seem to help much in
428 model specification. Besides, even the best forecast
429 combinations in the TvDM experiment do not produce
430 better results than the best forecasting model, so
431 searching for a good specification seems at this
432 point a more interesting question. However, paying
433 attention to the number of cases where RMSE is
434 reduced relative to the linear AR model, the best
435 combination is NN+STAR, using two nonlinear mod-
436 els. It would be interesting to see whether this turns
437 out to be a robust result.