XII Seminário Anual de Metas para a Inflação
13 e 14 de maio de 2010 – Rio de Janeiro

Forecasting Brazilian inflation using a large data set (preliminary version)

Francisco Marcos R. Figueiredo

BANCO CENTRAL DO BRASIL

Introduction

- Since policy decisions are taken based on the future inflation, forecasting inflation is a prime activity in Central Banks.
- Central banks monitor hundreds or even thousands of variables.
  - Central Bank of Brazil: Economic Indicators
- Traditional models for forecasting inflation: Short-run Phillips curve, VAR and its extensions (SVAR and BVAR)

The above models do not exploit the data-rich environment

- Stock and Watson (2006) – Forecasting with large datasets
  - Combining information: Factor and PLS models
  - Combining forecasts: “Traditional” forecast combination, BMA and Bagging

Objectives

- The objective is to verify if using large data set it is possible to obtain models that outperform the models commonly used by the monetary authorities for forecasting inflation
- Methods: Factor analysis by principal components and Partial Least Squares

Data-rich methodology I: Factor model

- Basic idea: Combining information of a large number of variables into few representative factors.
- Literature:
  - Sargent and Sims (1977)
  - APT model, core inflation indicators, money index and human development index and reaction functions
- Advantages
  - Factor modelers can remain agnostic about structure of the economy
  - Cope with many variables without having degree of freedom problems

Literature on forecasting using factor analysis

– Eickmeier and Ziegler (2008): 47 papers for more than 20 countries

Table 3.1 Summary of factor model results for forecasting inflation: RMSFE relative to autoregressive models

Papers	Country	Variable	Number of series	Forecast horizon							
Monthly data				**1**	**3**	**6**	**9**	**12**	**24**		
Moser, Rumler & Scharler (2007)	Austria	HICP	179	-	-	-	-	0.44	-		
Aguirre & Céspedes (2004)	Chile	CPI	306	-	0.95	1.05	0.61	0.56	-		
Marcellino *et al.* (2003)	Euro Area	CPI	401*	-	1.04	0.94	-	0.57	-		
Camacho & Sancho (2003)	Spain	CPI	1133	-	0.66	0.41	-	0.33	-		
Artis, Banerjee and Marcellino (2005)	UK	CPI	81	-	-	0.6	-	0.43	0.41		
Zaher (2005)	UK	CPI	167	-	-	-	-	0.65	-		
Stock and Watson (2002)	US	CPI	215	-	-	0.71	-	0.64	0.61		
Gavin and Kliesen (2006)	US	CPI	157	-	0.92	-	-	0.94	0.98		
Quarterly data				**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Gosselin & Tkacz (2001)	Canada	CPI	444	-	-	-	-	0.61	-	-	-
Angelini, Henry and Mestre (2001)	Euro Area	HICP	278	0.82	0.53	0.66	0.69	-	-	-	0.74
Matheson (2006)	New Zealand	CPI	384**	0.86	0.97	0.85	1.04	1.06	1.08	1.09	0.92

* Balanced panel
** The authors use data reduction rules
Source: Papers referred above and Eickmeier & Ziegler (2006)

– Factor models outperform benchmark models

The factor model: specification

Assuming the variables can be represented by an approximate linear dynamic factor structure with *r* common factors

$$X_{it} = \lambda_i(L) f_t + e_{it}$$

X_{it} represents the observed value of explanatory variable *i* at time *t*
f_t is the *r x 1* vector of non-observable factors and
e_{it} is the idiosyncratic component.

The problem is to minimize the following non-linear objective function:

$$V(F,\Lambda) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{i=1}^{T} \left(X_{it} - \lambda'_i F_t \right)^2$$

The factor model: estimation

When ε_{it} is both serially correlated and weakly cross-sectionally correlated, Stock and Watson (2002) show that F_t can be estimated by the standard method of principal components

$$\hat{F} = X\hat{\Lambda} / N$$

$\hat{\Lambda}$ is equal to $N^{1/2}$ times the eigenvectors of the $N \times N$ matrix $X'X$ corresponding to its largest r eigenvalues.

- An estimated factor can be thought as a weighted average of the series in the dataset, where the weights can be either positive or negative and reflect how correlated each variable is with each factor.
- Factors are obtained in a sequential way, with the first factor explaining the most variation in the dataset, the second factor explaining the most variation not explained by the first factor, and so on.

The factor model: empirical issues

- Choosing the optimal number of factors
  - Rules of thumb, Forecast performance
  - Bai and Ng (2002): $IC = \ln(\hat{V}_r) + rg(T, N)$
- Data with different frequencies and missing values
- Choosing the "optimal" data size
  - Initially: the larger, the better
  - Bai and Ng (2006) show that extracting factor does not always yield better forecasting performance
  - Targeting the predictors: leading indicators and forecasting ability

Data-rich methodology II: Partial Least Squares (PLS)

- Econometric technique developed by Wold (1966) is popular among chemical engineers and chemometricians
- PC factors are obtained taking into account only the predictor variables, whereas in PLS, the relationship between the predictors and the variable to be forecasted is considered for constructing the factors.
- PLS searches for a set of components that performs simultaneous decomposition of X and y with the constraint that these components explain as much as possible of the covariance between X and y .
- Few examples in forecasting macroeconomic variables so far:
- Lin and Tsay (2006), Groen and Kapetanios (2008) and Eickmeier and Ng (2009)

Algorithm for the Partial Least Squares (PLS)

Helland (1990), Groen and Kapetanios (2008) and Eickmeier and Ng (2009)

1) Set $u_t = y_t$ and $v_{i,t} = x_{i,t}$, *i* = 1, … N. Set *j* = 1;

2) Determine *N x 1* vector of loading $w_j = (w_{1j} \cdots w_{Nj})$ by computing individual covariances: $w_{ij} = cov(u_t, v_{it})$, *i* = 1, … N. Construct the *j-th* PLS factor by taking the linear combination given by $w_j' v_t$ and denote this factor by $f_{j,t}$;

3) Regress u_t and $v_{i,t}$, *i* = 1, … N on $f_{j,t}$. Denote the residuals of these regressions by $\tilde{u}_t$ and $\tilde{v}_{i,t}$ respectively and

4) If *j* = *k* stop, else set $u_t = \tilde{u}_t, v_{i,t} = \tilde{v}_{i,t}$ *i* = 1, … N and *j* = *j+1* and go to step 2.

Estimation and forecasting framework

Principal Component Factor Model (PC) and Partial Least Square Model (PLS):

Initial data set → Few factors → Inflation forecasts

Targeted Principal Component Factor Model (TPC):

Initial data set → “Targeted” predictors → Few factors → Inflation forecasts

Forecasting framework

Dynamic estimation: direct forecasts (Clements and Hendry, 1996)

$$y_{t+h}^{h} = \mu + \alpha(L) y_t + \beta(L) Z_t + \varepsilon_{t+h}^{h}$$

The dependent variables are headline IPCA inflation and market prices inflation

$$y_{t+h}^{h} = \frac{\ln(IPCA_{t+h} / IPCA_t)}{h}$$

Out-of-sample forecasts: recursive and rolling estimation

The factor models were estimated for the balanced panel with $1 \leq r \leq 6$ (number of factors), $1 \leq m \leq 4$ (number of the lags for the factors) and $0 \leq p \leq 6$ (number of the lags for inflation).

The Brazilian data and forecast horizon

- The initial dataset for Brazil contains 368 monthly series over the sample period of January 1995 to July 2009.
- Treatment: logarithms, unit root tests, seasonal adjustment, zero mean and unit variance.
- Forecast horizon: January 2001 to July 2009.
- Targeting the predictors through Granger-causality tests.

The datasets

Table 6.1 - Variables employed in factors estimation

Sectors	Number of variables
Monetary Aggregates	13
Credit	12
Interest rates	9
Fiscal variables	25
Exchange rates	22
Price indices	81
Industrial production	47
Production and inventories	14
Capacity utilization	3
Consumption and sales	24
Employment and working hours	32
Wages and payroll	11
Default	6
External sector	49
International	15
Miscellaneous	5
Overall	**368**

Table 6.2 - Number of targeted predictors

Horizon	Headline	Market Prices
Overall	**368**	**368**
1-step-ahead	94	108
3-step-ahead	109	110
6-step-ahead	115	108
9-step-ahead	120	128
12-step-ahead	116	143

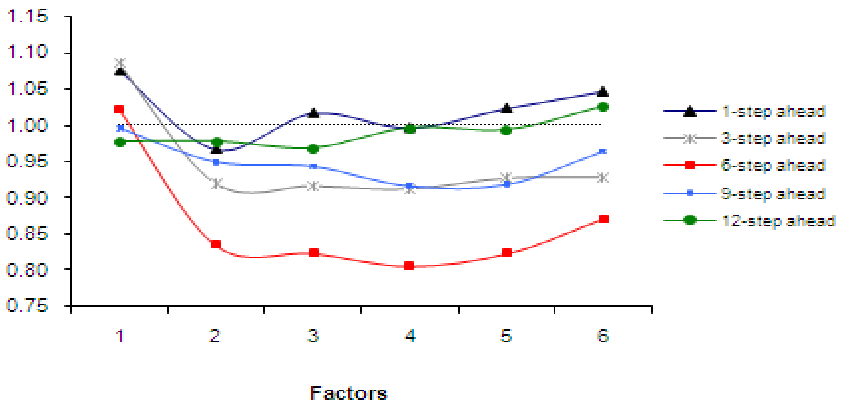
Models

- Approaches
  - Factor model with principal components (PC)
  - Factor model with principal components and targeted variables (TPC)
  - Partial least Squares (PLS)
- Estimation
  - Recursive regression
  - Rolling regression
- Variable
  - Headline inflation
  - Market price inflation

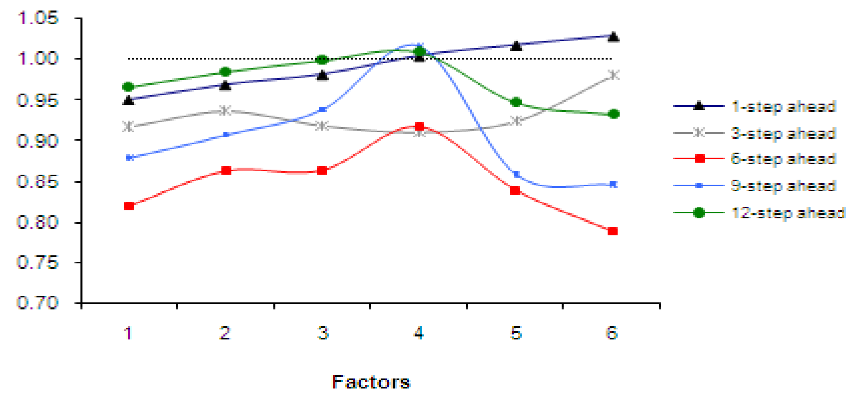
Out-of-sample forecasts: headline inflation – recursive regressions

Figure - Relative RMSE for headline inflation 2001-2009 recursive median models

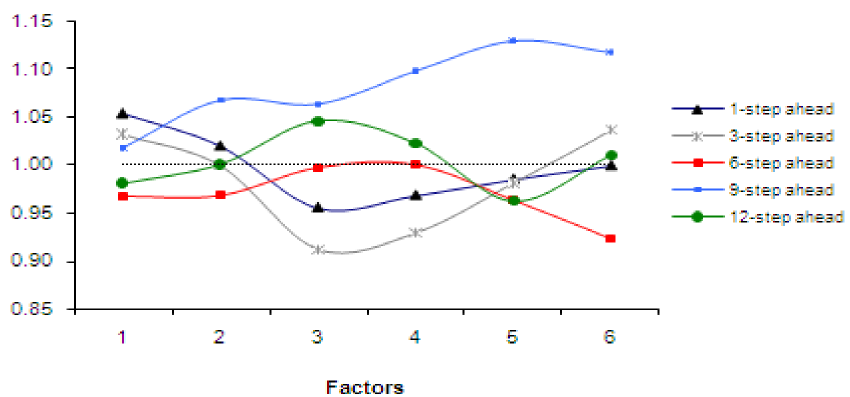
Principal component



Targeted principal component



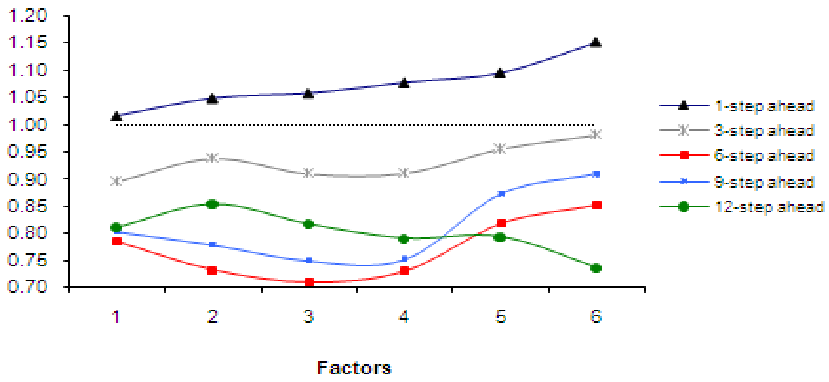
Partial least squares



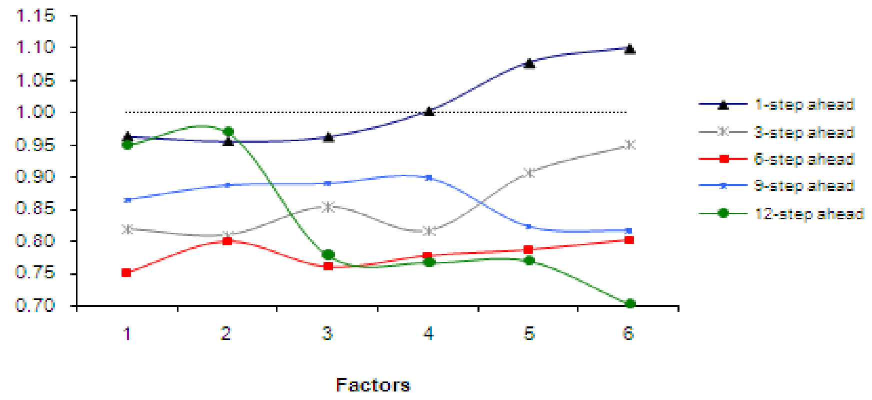
Out-of-sample forecasts: headline inflation – rolling regressions

Figure - Relative RMSE for headline inflation 2001-2009 rolling median models

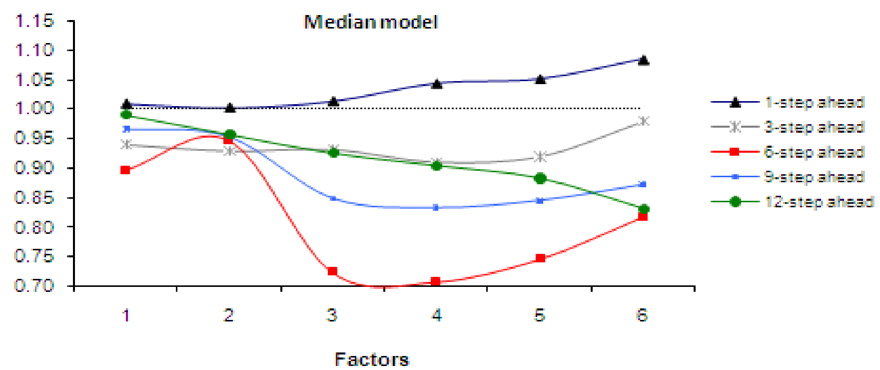
Principal component



Targeted principal component



Partial least squares



Out-of-sample forecasts: headline inflation

Figure 7.6 Relative RMSFE for headline recursive models

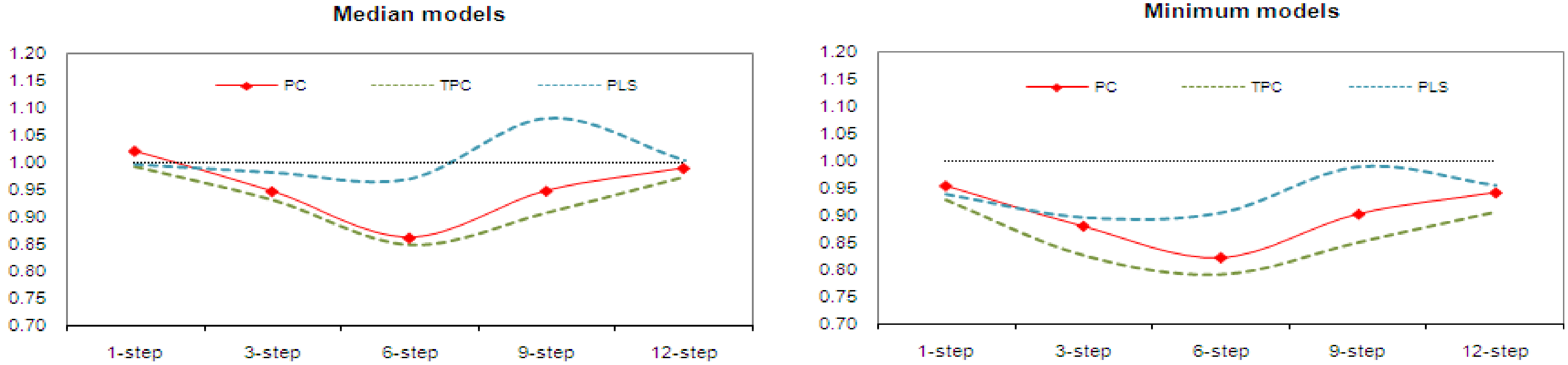
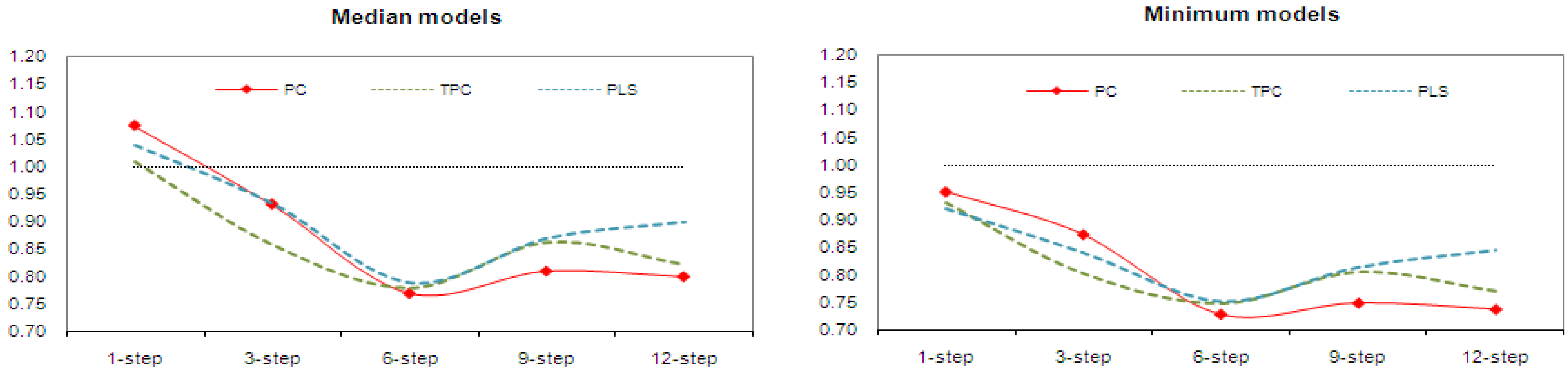


Figure 7.7 Relative RMSFE for headline rolling models



VAR and BVAR models

Table 5.1 Specifications of VAR models used by Central Bank of Brazil

Endogenous variables	**VAR models**			
	Unrestricted		**Bayesian**	
	1	2	3	4
Real interest rate	x			
Nominal interest rate		x	x	x
Money stock		x	x	x
Industrial output		x	x	x
Nominal exchange rate	x	x	x	x
Regulated price	x	x	x	x
Market price	x	x	x	x
Deterministic components				
Constant	x	x	x	x
Three trend dummies	x	x	x	x
Seasonal dummies	x	x		x
Lags	2	6	6	6

Source: Inflation Report, Central Bank of Brazil, June 2004

Out-of-sample forecasts: Diebold-Mariano Test

Table 6.3 - Comparing the predictive accuracy of the models

Horizon	Model	Var 1	Var 2	Bvar 1	Bvar 2	PC	TPC	PLS
1-step ahead	**Var 1**	-	**2.360**	**-2.178**	-0.851	-1.446	*-1.830*	-0.112
	Var 2		-	**-2.988**	**-2.749**	**-2.599**	**-2.727**	*-1.751*
	Bvar 1			-	*1.747*	-0.374	-1.089	1.219
	Bvar 2				-	-1.224	*-1.734*	0.211
	PC					-	-1.245	**2.500**
	TPC						-	**3.508**
3-step ahead	**Var 1**	-	**2.463**	**-2.761**	**-2.323**	**-2.309**	**-2.352**	-0.859
	Var 2		-	**-3.198**	**-3.196**	**-2.914**	**-2.793**	*-1.938*
	Bvar 1			-	0.041	*-1.701*	*-1.738*	-0.003
	Bvar 2				-	*-1.807*	*-1.813*	-0.014
	PC					-	-0.929	**2.167**
	TPC						-	**2.501**
6-step ahead	**Var 1**	-	1.604	*-1.938*	**-2.437**	**-2.111**	**-2.576**	-0.801
	Var 2		-	**-2.424**	**-2.678**	**-2.505**	**-2.755**	-1.368
	Bvar 1			-	-0.514	*-1.922*	**-2.503**	-0.397
	Bvar 2				-	*-1.765*	**-2.359**	-0.271
	PC					-	**-2.969**	**2.525**
	TPC						-	**4.771**
9-step ahead	**Var 1**	-	**2.308**	-1.301	*-1.888*	*-1.828*	**-2.720**	-0.051
	Var 2		-	**-3.403**	**-3.172**	**-2.544**	**-3.292**	-0.674
	Bvar 1			-	-0.183	*-1.722*	**-2.607**	0.138
	Bvar 2				-	*-1.699*	**-2.602**	0.158
	PC					-	-1.271	**3.501**
	TPC						-	**3.711**
12-step ahead	**Var 1**	-	**3.026**	*-1.711*	-1.397	*-1.801*	**-3.210**	-1.311
	Var 2		-	**-3.523**	**-3.524**	**-2.838**	**-3.999**	**-2.225**
	Bvar 1			-	0.743	-1.661	**-3.169**	-1.176
	Bvar 2				-	*-1.710*	**-3.202**	-1.207
	PC					-	**-2.137**	0.491
	TPC						-	**2.332**

Diebold-Mariano test statistic. Bold and italic figures indicate rejection of the null of equal predictive accuracy at 5% and 10% significance levels respectively.

Positive (negative) values mean that the model in the row (column) presents a higher predictive accuracy than that of the model given by the column (row). Bold figures (italic figures) indicate that the statistic is significant at 5% (10%) significance level.

Out-of-sample forecasts: Encompassing test

Table 6.4 - Forecast encompassing test: p-values for the null hypothesis of no predictive power

		Var 1	Var 2	Bvar 1	Bvar 2	PC	TPC	PLS
1-step ahead	Var 1		0.760	0.000	0.000	0.000	0.001	0.001
	Var 2	0.000		0.000	0.000	0.000	0.000	0.000
	Bvar 1	0.984	0.309		0.051	0.009	0.011	0.036
	Bvar 2	0.548	0.039	0.092		0.001	0.001	0.003
	PC	0.448	0.781	0.930	0.348		0.020	0.499
	TPC	0.901	0.795	0.939	0.403	0.243		0.268
	PLS	0.111	0.703	0.181	0.018	0.000	0.000	
3-step ahead	Var 1		0.484	0.020	0.027	0.004	0.000	0.003
	Var 2	0.002		0.000	0.001	0.002	0.000	0.001
	Bvar 1	0.659	0.856		0.079	0.009	0.000	0.007
	Bvar 2	0.596	0.300	0.619		0.016	0.001	0.012
	PC	0.998	0.705	0.632	0.904		0.001	0.171
	TPC	0.630	0.478	0.174	0.497	0.181		0.797
	PLS	0.718	0.927	0.870	0.617	0.005	0.001	
6-step ahead	Var 1		0.484	0.020	0.027	0.004	0.000	0.003
	Var 2	0.002		0.000	0.001	0.002	0.000	0.001
	Bvar 1	0.659	0.856		0.079	0.009	0.000	0.007
	Bvar 2	0.596	0.300	0.619		0.016	0.001	0.012
	PC	0.998	0.705	0.632	0.904		0.001	0.171
	TPC	0.630	0.478	0.174	0.497	0.181		0.797
	PLS	0.718	0.927	0.870	0.617	0.005	0.001	
9-step ahead	Var 1		0.861	0.203	0.167	0.022	0.002	0.058
	Var 2	0.000		0.000	0.000	0.003	0.000	0.005
	Bvar 1	0.276	0.772		0.223	0.027	0.001	0.067
	Bvar 2	0.715	0.323	0.745		0.037	0.001	0.085
	PC	0.702	0.987	0.840	0.866		0.000	0.738
	TPC	0.977	0.822	0.590	0.676	0.105		0.817
	PLS	0.675	0.980	0.890	0.786	0.042	0.000	
12-step ahead	Var 1		0.088	0.430	0.437	0.043	0.002	0.015
	Var 2	0.000		0.000	0.000	0.001	0.000	0.000
	Bvar 1	0.424	0.061		0.726	0.049	0.000	0.009
	Bvar 2	0.350	0.023	0.443		0.053	0.000	0.010
	PC	0.508	0.741	0.664	0.691		0.001	0.043
	TPC	0.459	0.063	0.120	0.155	0.563		0.854
	PLS	0.758	0.462	0.991	0.983	0.398	0.001	

P-values for the null hypothesis of no predictive power of model in the column with respect to the model in the row.

P-values for the null hypothesis of no predictive power of model in the column with respect to the model in the row.

Concluding remarks and research agenda

- Findings:
  - ✓ Best relative performance for 6-step ahead forecast
  - ✓ Rolling regression models outperform recursive models
  - ✓ PLS performance is poor and TPC display the best results
- Research agenda:
  - ✓ Factor model and PLS: other approaches and algorithms
  - ✓ Checking the robusteness of the results over different samples
  - ✓ Quarterly data, combining frequencies and missing values
  - ✓ Other methods: BMA and bagging