Journal of Forecasting
J. Forecast. (2010)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/for.1182

Forecasting Regional House Price Inflation: A Comparison between Dynamic Factor Models and Vector Autoregressive Models

SONALI DAS,[1] RANGAN GUPTA[2]* AND ALAIN KABUNDI[3]
[1] *Logistics and Quantitative Methods, CSIR Built Environment, Pretoria, South Africa*
[2] *Department of Economics, University of Pretoria, Pretoria, South Africa*
[3] *Department of Economics, University of Johannesburg, Johannesburg, South Africa*

ABSTRACT

This paper uses the dynamic factor model framework, which accommodates a large cross-section of macroeconomic time series, for forecasting regional house price inflation. In this study, we forecast house price inflation for five metropolitan areas of South Africa using principal components obtained from 282 quarterly macroeconomic time series in the period 1980:1 to 2006:4. The results, based on the root mean square errors of one to four quarters ahead out-of-sample forecasts over the period 2001:1 to 2006:4 indicate that, in the majority of the cases, the Dynamic Factor Model statistically outperforms the vector autoregressive models, using both the classical and the Bayesian treatments. We also consider spatial and non-spatial specifications. Our results indicate that macroeconomic fundamentals in forecasting house price inflation are important.

KEY WORDS Bayesian models; forecast accuracy; spatial and non-spatial models

INTRODUCTION

This paper investigates whether the wealth of information contained in the dynamic factor model (DFM) framework, developed by Forni *et al.* (2005), can be useful in forecasting regional house price inflation. To illustrate, we use the DFM to predict house price inflation, defined as the percentage change in house prices, in five metropolitan areas of South Africa, namely Cape Town, Durban, Johannesburg, Port Elizabeth and Pretoria, using quarterly data over the period 1980:1 to 2006:4. The panel data comprise 282 quarterly series for the South African economy, a set of global variables such as commodity industrial inputs price index and crude oil prices, and time series of major trading

*Correspondence to: Rangan Gupta, Department of Economics, University of Pretoria, Pretoria 0002, South Africa. E-mail: Rangan.Gupta@up.ac.za

partners, namely Germany, the UK, and the USA. The forecast performance of the DFM is evaluated in terms of the root mean square error (RMSE), by comparing its performance with spatial Bayesian vector autoregressive (SBVAR) models that weighs in the influence of neighbors on the determination of house price inflation of a particular region, and also to the non-spatial unrestricted classical vector autoregressive (VAR) model and Bayesian vector autoregressive (BVAR) models using the Minnesota prior. All the alternative models are estimated based on only the house price inflation series.

The importance of predicting house price inflation is highlighted by recent studies in which it is concluded that asset prices help to forecast both inflation and output (Forni *et al.*, 2003; Stock and Watson, 2003). Since a large amount of individual wealth is embedded in houses, house prices are important in signaling inflation. Gupta and Das (2008) point out that, in South Africa, housing inflation and consumer price index (CPI) inflation tend to move together. As such, models that forecast house price inflation can give policy makers an idea about the direction of CPI inflation in the future, and hence can provide a better control for designing appropriate policies. The reason for using regional data is to account for possible heterogeneity and segmentation that might exist in the housing market. Herein also comes the justification of modeling house prices separately based on the size of the house.

The rationale behind using a data-rich DFM to forecast house price inflation emanates from the fact that a large number of economic variables help in predicting housing price growth (Cho, 1996; Abraham and Hendershott, 1996; Johnes and Hyclak, 1999; Rapach and Strauss, 2007, 2009). For instance, income, interest rates, construction costs, labor market variables, stock prices, industrial production, and consumer confidence index—which are included in the DFM—are potential predictors. In addition, given that movements in the housing market are likely to play an important role in the business cycle, not only because housing investment is a very volatile component of demand (Bernanke and Gertler, 1995), but also because changes in house prices tend to have important wealth effects on consumption (International Monetary Fund, 2000) and investment (Topel and Rosen, 1988), the importance of forecasting house price inflation is vital, since the housing sector serves as a leading indicator of the real sector of the economy.

To the best of our knowledge, this is the first attempt to compare the forecasting performances of a DFM with spatial and non-spatial econometric models in terms of predicting regional house price inflation. Rapach and Strauss (2007, 2009) used autoregressive distributed lag (ARDL) models containing, respectively, 25 and 30 variables to forecast real housing price growth for the individual states of the Federal Reserve's Eighth District and the 20 largest US states. Given the difficulty in determining a priori the particular variables that are most important for forecasting real housing price growth, the authors used various methods to combine the individual ARDL model forecasts, which in turn resulted in better forecasting of real housing price growth compared to the individual ARDL models. Vargas-Silva (2008) point to the importance of using as many as 120 monthly series to analyze the impact of monetary policy actions on the housing sector of four different regions of the USA based on a factor-augmented VAR (FAVAR) model. The author indicates that the housing market tends to be negatively affected by positive interest rate shocks but, more importantly, marked heterogeneity in the responses of the housing market variables for the four regions were depicted, with the Southern region driving the aggregate US housing market. The remainder of the paper is organized as follows: the second and third sections, respectively, lay out the DFM and outline the basics of the VAR, the Minnesota-type BVARs, and the SBVARs based on the first-order spatial contiguity (FOSC) and the random walk averaging (RWA) priors developed by LeSage and Pan (1995) and LeSage and Krivelyova (1999), respectively. The fourth section discusses the data used

DOI: 10.1002/for

to estimate the DFM, while the fifth section reports the results from the forecasting exercise. The sixth section concludes.

DYNAMIC FACTOR MODEL (DFM)

This study uses the generalized dynamic factor model (DFM) developed by Forni *et al.* (2005) to extract common components between macroeconomics series, which are then used to forecast metropolitan house price inflation for the South African housing market. In the VAR models, since all variables are used in forecasting, the number of parameters to be estimated depends on the number of variables n. With such a large information set, n, the estimation of a large number of parameters leads to a curse of dimensionality problem. The generalized DFM expresses individual times series as the sum of two unobserved components: a common component driven by a small number of common factors and an idiosyncratic component, which are specific to each variable. Forni *et al.* (2005) demonstrated that when the number of factors is small relative to the number of variables and the panel is heterogeneous, the factors can be recovered from present and past observations.

Consider an $n \times 1$ covariance stationary process $Y_t = (y_{1t}, \ldots, y_{nt})'$. Define X_t as the standardized version of Y_t. The generalized DFM of by Forni *et al.* (2005) is given by

$$X_t = B(L) f_t + \xi_t = \Lambda F_t + \xi_t \tag{1}$$

where f_t is a $q \times 1$ vector of dynamic factors, while F_t is an $r \times 1$ vector of static factors, with $r = q(s + 1)$, $B(L) = B_0 + B_1 L + \ldots + B_s L^s$ is an $n \times q$ factor loadings matrix polynomial of order s and Λ is the factor loadings matrix related to static factors, ξ_t is the $n \times 1$ vector of idiosyncratic components. Note that L denotes the lag operator.

The generalized DFM is a weighted version of the static principal components estimator of Stock and Watson (2002), which exploits information of leads and lags of variables where time series are converted to the frequency domain. However, dynamic principal component analysis (PCA) is a two-sided filter. This causes a problem at the end of the sample, making it difficult to estimate and forecast the common component since no future observations are available. The generalized DFM solves this problem by proposing a two-step approach. In the first step, it relies on the dynamic approach in the estimation of the covariance matrices of the common and idiosyncratic component (at all leads and lags) through an inverse Fourier transform of the spectral density matrices. It involves estimating the eigenvalues and eigenvectors decomposition of the spectral density matrix for X_t, $\hat{\Sigma}(\theta)$ with rank q, corresponding to the q largest eigenvalues. For each frequency, $-\pi < \theta < \pi$, the spectral density matrix of X_t can be decomposed into the spectral densities of the common and the idiosyncratic component, $\Sigma(\theta) = \Sigma_\chi(\theta) + \Sigma_\xi(\theta)$. Hence the estimated spectral density matrix of common component $\hat{\Sigma}_\chi(\theta)$ can be constructed. In the second step, this information is used to compute r linear combinations of X_t that maximizes the contemporaneous covariance matrices estimated explained by the common factors, estimated in the first step. Further, it reduces the idiosyncratic noise in the common factor space to a minimum, by selecting the variables with the highest common/idiosyncratic variance ratio. This step is a one-sided approach and is only used to estimate and forecast the common component.

For forecasting purposes, we adopt the method of Boivin and Ng (2005). The forecasting equation is as follows:

DOI: 10.1002/for

$$\hat{y}_{t+h} = \hat{\chi}_{t+h} + \hat{\phi}(L)\hat{\xi}_t \tag{2}$$

where $\hat{\chi}_{t+h}$ is obtained by artificially projecting χ_{t+h} on the estimated dynamic factor, $\hat{F}_t$ obtained from (1), such that $\hat{\chi}_{t+h} = \hat{\Gamma}_\chi(k)Z(Z'\hat{\Sigma}Z)^{-1}Z'X_t$, Z is the r generalized eigenvectors of $\hat{\Gamma}_\chi(k)$ with respect to $\hat{\Gamma}_\xi(k)$ under normalization $Z'\hat{\Gamma}_\chi(0)Z = 1$, and $\hat{\Gamma}_\chi(k)$ and $\hat{\Gamma}_\xi(k)$ are covariance matrices of common and idiosyncratic components at different leads and lags. Since $\hat{F}_t = Z'X_t$, then $\hat{\chi}_{t+h} = \hat{\Gamma}_\chi(\theta)Z(Z'\hat{\Sigma}Z)^{-1}\hat{F}_t$.

VECTOR AUTOREGRESSIVE (VAR) MODELS

In this study, the generalized DFM is our benchmark model. To evaluate the forecasting performance of the DFM we consider alternative models: in our case, the unrestricted classical VAR, BVARs based on the Minnesota prior, and the SBVARs based on the FOSC and RWA priors. This section outlines the basics of the above-mentioned competing models.

Classical VARs

The VAR model, as suggested by Sims (1980), can be written as follows:

$$y_t = A_0 + A(L)\,y_t + \varepsilon_t \tag{3}$$

where y is an $n \times 1$ vector of variables being forecast; $A(L)$ is an $n \times n$ polynomial matrix in the backshift operator L with lag length p, i.e., $A(L) = A_1L + A_2L^2 + \ldots + A_pL^p$; A_0 is an $n \times 1$ vector of constant terms; and ε is an $n \times 1$ vector of error terms. In our case, we assume that $\varepsilon_t \approx N(0, \sigma^2 I_n)$, where I_n is an $n \times n$ identity matrix. The VAR model is estimated based on ordinary least squares (OLS) and forecasting is straightforward (See Hamilton, 1994, ch. 11).

Non-spatial Bayesian VARs

The BVAR model, on the other hand, as described in Litterman (1981), Doan *et al.* (1984), Todd (1984), Litterman (1986), and Spencer (1993), imposes priors on the coefficients of the VAR model. Besides a diffuse prior on the constants, the means of the prior, popularly called the Minnesota prior, take the following form:

$$\beta_i \sim N\left(1, \sigma_{\beta_i}^2\right) \text{ and } \beta_j \sim N\left(0, \sigma_{\beta_j}^2\right) \tag{4}$$

where β_i denotes the coefficients associated with the lagged dependent variables in each equation of the VAR, while β_j represents any other coefficient. The specification of the standard deviation of the distribution of the prior imposed on variable j in equation i at lag m, for all i, j and m, denoted by $S_1(i, j, m)$, is specified as follows:

$$S_1(i, j, m) = [w \times g(m) \times F(i, j)]\frac{\hat{\sigma}_i}{\hat{\sigma}_j} \tag{5}$$

with $F(i, j) = 1$, if $i = j$ and k_{ij} otherwise, with $(0 \le k_{ij} \le 1)$; $g(m) = m^{-d}$, $d > 0$. Note that $\hat{\sigma}_i$ is the estimated standard error of the univariate autoregression for variable i. The ratio $\hat{\sigma}_i/\hat{\sigma}_j$ scales the

DOI: 10.1002/for

variables to account for differences in the units of measurement, and hence causes specification of the prior without consideration of the magnitudes of the variables. The term w indicates the overall tightness and is also the standard deviation on the first own lag, with the prior getting tighter as we reduce the value. The parameter $g(m)$ measures the tightness on lag m with respect to lag 1, and is assumed to have a harmonic shape with a decay factor of d, which tightens the prior on increasing lags. The parameter $F(i, j)$ represents the tightness of variable j in equation i relative to variable i, and by increasing the interaction, i.e., the value of k_{ij}, we can loosen the prior. Note that the overall tightness (w) and the lag decay (d) hyperparameters used in the standard Minnesota prior have values of 0.1 and 1.0, respectively, while $k_{ij} = 0.5$ implies a weighting matrix (F) with 1.0 on the diagonals and 0.5 as the off-diagonal elements.

Spatial Bayesian VARs

Given that the Minnesota prior treats all variables in the VAR, except for the first own lag of the dependent, in an identical manner, several attempts have been made to alter this fact. Usually, this has boiled down to increasing the value of the overall tightness (w) hyperparameter from 0.10 to 0.20, so that the larger value of w can allow for more influence from other variables in the model. In addition, as proposed by Dua and Ray (1995), we also try out a prior that is even more loose, specifically with $w = 0.30$ and $d = 0.50$. Alternatively, LeSage and Pan (1995) have suggested the construction of the weight matrix based on the FOSC, which implies the creation of a non-symmetric F matrix that emphasizes the importance of the variables from the neighboring states/provinces more than those of the non-neighboring states/provinces. Lesage and Pan (1995) suggest the use of a value of unity on the diagonal elements of the weight matrix, as in the Minnesota prior, as well as in place(s) that correspond to the variable(s) from other state(s)/province(s) with which the specific state in consideration has common order(s). However, for the elements in the F matrix that corresponds to variable(s) from state(s)/province(s) that are not immediate neighbor(s), Lesage and an (1995) propose a value of 0.1. (See the Appendix for further details regarding the construction of the F matrix based on FOSC.)

In addition to the FOSC-based prior, LeSage and Krivelyova (1999) have also put forth another approach to remedy the equal treatment nature of the Minnesota prior by the RWA prior. This involves both the prior means and the variances based on a distinction made between important variables (like house price inflation of neighboring metropolitan area(s)), and unimportant variables (like house price inflation of non-neighboring metropolitan area(s)), in each equation of the VAR model. To understand the motivation behind the design of the prior means, consider the weight matrix F for the VAR consisting of house price inflation of the five metropolitan areas. Retaining the ordering of the five metropolitan areas as outlined in the FOSC prior, the weight matrix contains values of unity in positions associated with the house price inflation(s) of neighboring metropolitan area(s), i.e., for important variables in each equation of the VAR model, while zero values are assigned to the unimportant variables, i.e., house price inflation of non-neighboring metropolitan area(s), with neighbors and non-neighbors identified as discussed under the FOSC prior. As with the Minnesota prior, we continue to have a value of one on the main diagonal of the F matrix, which is then standardized to a matrix C, so that the rows sum to unity and we can consider the random walk with drift, which averages over the important variables in each equation i of the VAR. (See the Appendix for further details on the design of the F and C matrices under RWA prior. The prior on the standard deviations for the RWA-based model has also been outlined in the Appendix.)

DOI: 10.1002/for

Estimation of BVARs

Finally, the BVARs and the SBVARs, based on the FOSC and the RWA priors, are estimated using Theil's (1971) mixed estimation technique (See Hamilton, 1994, pp. 360–362). In each equation of the different types of VARs, there are 41 parameters including the constants, given the fact that the model is estimated with eight lags of each variable, which are essentially the house price inflation (percentage change in house prices) of the five metropolitan areas. All data on house price inflation are seasonally adjusted, before being converted to house price inflation, in order to (*inter alia*) address the fact that, as pointed out by Hamilton (1994, p. 362), the Minnesota-type priors are not well suited for seasonal data. The choice of eight lags is based on the unanimity of the sequential modified LR test statistic, Akaike information criterion (AIC) and the final prediction error (FPE) criterion.

The five-variable VAR, BVAR and SBVAR models for an initial prior are estimated for the period of 1980:1 to 2000:4 and then forecast from 2001:1 through to 2006:4. Since we use eight lags, the initial eight quarters of the sample 1980:1 to 1981:4 are used to feed the lags. We generate dynamic forecasts, as would naturally be achieved in actual forecasting practice. The models are re-estimated each quarter over the out-of-sample forecast horizon in order to update the estimate of the coefficients, before producing the four-quarters-ahead forecasts. This iterative estimation and four-steps-ahead forecast procedure was carried out for 24 quarters, with the first forecast beginning in 2001:1. This experiment produced a total of 24 one-quarter-ahead forecasts, 24 two-quarters-ahead forecasts, and so on, up to 24 four-steps-ahead forecasts. The RMSEs for the 24 quarter 1 through quarter 4 forecasts, for the period 2001:1 to 2006:4, are then calculated and compared for the house price inflation of the five metropolitan areas obtained from that of the generalized DFM. Note that if A_{t+n} denotes the actual value of a specific variable in period $t + n$ and ${}_tF_{t+n}$ is the forecast made in period t for $t + n$, the RMSE statistic can be defined as

$$\sqrt{\frac{1}{N}\sum (A_{t+n} - {}_tF_{t+n})^2} \times 100$$

For $n = 1$, the summation runs from 2001:1 to 2006:4, and for $n = 2$ the same covers the period of 2001:2 to 2006:4, and so on.

DATA

We study the South African house price data empirically. As in Burger and Van Rensburg (2008) and Gupta and Das (2008), we do not consider the residential market in general; rather, we subdivide the market in terms of size and price of the houses. Specifically, we use the ABSA (one of the leading South African private banks) house price index, which distinguishes between three price categories, expressed in the domestic currency rand (R), as luxury houses (R2.6 million to R9.5 million), middle-segment houses (R226,000 to R2.6 million) and affordable houses (R226,000 and below). Further, the middle-segment category has three subcategories based on size (measured by square meters of house): small (80–140 m^2), medium (141–220 m^2) and large (221–400 m^2). Given that regional house price inflation data are only available for the middle-segment houses, we restrict our study to this category. Although ABSA reports data for both metropolitan and non-metropolitan areas, the availability is limited and lacks clarity regarding the area of coverage, especially for the rural areas. We thus limit our analysis to the five major metropolitan areas of South Africa.

DOI: 10.1002/for

Besides the 15 house price series (five metropolitan areas for each of the three middle-segment houses), the dataset contains 267 quarterly series of South Africa, ranging from real, nominal, and financial sectors; intangible variables, such as confidence indices and survey variables, and additionally to national variables; we also use a set of global variables such as commodity industrial inputs price index and crude oil prices. The data comprise series of major trading partners such as Germany, the UK and the USA. All series are seasonally adjusted. The more powerful Dickey–Fuller generalized least squares (DF-GLS) test of Elliott *et al.* (1996) is used to assess the degree of integration of all series. All non-stationary series are made stationary through differencing. The Schwarz information criterion is used to select the appropriate lag length so that no serial correction is left in the stochastic error term. Where there were doubts about the presence of unit root, the KPSS test (Kwiatowski *et al.*, 1992), with the null hypothesis of stationarity, was applied. All series are standardized to have a mean of zero and a constant variance. The in-sample period contains data from 1980:1 to 2000:4, while the out-of-sample set is 2001:1 to 2006:4.

There are various statistical approaches in determining the number of factors in the DFM. The Bai and Ng (2002) approach suggests five static factors in our dataset, while the Bai and Ng (2007) approach suggests two dynamic factors. The approach of Forni *et al.* (2000) also suggests two dynamic factors, with the first two dynamic principal components explaining approximately 99% of the variation.

EVALUATION OF FORECAST ACCURACY

To evaluate the accuracy of forecasts generated by the DFM, we compare its performance with the alternative models using the same statistic, namely the RMSE. Given that there are seven alternative models, we use a parsimonious approach while reporting the results in Tables I–III.[1] We compare each of the one- to four-quarters-ahead forecasts generated by the DFM with those from a specific-type of VAR model that performs the best, in terms of average RMSEs[2] for one- to four-quarters-ahead forecasts, within the category of VAR models for a specific region, under a particular category of housing, over the out-of-sample horizon of 2001:1 to 2006:4.[3] Note the values for these hyperparameters are based on the ranges suggested by LeSage (1999). The main observations can be summarized as follows:

- *Large middle-segment houses*. From Table I we observe that for the category of large middle-segment houses the DFM, barring the cases of third- and fourth-quarter-ahead forecasts for Port Elizabeth, and first-, third- and fourth-quarter-ahead forecasts for Cape Town, outperforms the best-performing model within the VAR category. Amongst the alternative VARs, the SBVAR model based on the FOSC prior is the standout performer for all the metropolitan areas, except

[1] In Tables I–III, the metropolitan areas have been abbreviated to ECAP, JOBU, KWAZ, PRET, and WCAP for Eastern Cape (Port Elizabeth), Johannesburg, KwaZulu Natal (Durban Unicity), Pretoria, and Western Cape (Cape Town), respectively.

[2] The decision to use average RMSEs for choosing the best-performing VAR model within this category is standard practice. For two recent examples, refer to Liu *et al.* (2009a,b). However, all the other forecasting results for the alternative types of VAR models will be made available upon request.

[3] Note that for the SBVAR model based on the RWA prior that did best amongst other SBVAR models consistently for all house sizes and majority of the metropolitan areas had the following values of the hyperparameters: $\sigma_c = 0.3$, $\eta = 8$, and $\rho = 1$ (refer to the Appendix for further details).

DOI: 10.1002/for

Table I. RMSEs for large middle-segment houses (2001:1 to 2006:4)

Region	Model	Quarters ahead			
		1	2	3	4
ECAP	DFM	5.1070	5.2891	5.4843	5.7861
		(−3.4843***)	(−2.8597***)	(1.4047)	(1.4125)
	SBVAR1	11.9550	9.1444	4.1622	4.3177
JOBU	DFM	2.0938	2.8122	2.7223	2.5270
		(−1.9688**)	(−4.1425***)	(−3.3087***)	(−3.7419***)
	SBVAR1	4.1098	9.4896	7.0795	7.3525
KWAZ	DFM	4.8798	5.3617	5.2879	5.0944
		(−3.0842***)	(−2.8666***)	(−1.0528)	(−3.4439***)
	BVAR1	10.2062	8.4188	5.3277	9.9729
PRET	DFM	2.3077	3.4558	2.4556	2.5799
		(−1.9613**)	(−2.6433***)	(−3.1497***)	(−4.2843***)
	SBVAR1	4.9353	7.0118	6.4439	13.3315
WCAP	DFM	2.3846	2.3843	2.8635	2.9397
		(1.2997)	(−1.1664)	(1.3863)	(1.6584*)
	SBVAR1	2.0253	2.5750	2.4713	1.2203

Note: DFM, dynamic factor model; BVAR1, BVAR model based on the standard Minnesota prior with $w = 0.1$, $d = 1.0$; SBVAR1, spatial BVAR based on the first-order spatial contiguity (FOSC) prior. Numbers in parentheses represent the Diebold and Mariano (1995) tests statistic, with asterisks indicating significance at ***1%, **5% and *10% levels, respectively.

Table II. RMSEs for medium middle-segment houses (2001:1 to 2006:4)

Region	Model	Quarters ahead			
		1	2	3	4
ECAP	DFM	3.3023	4.0077	3.5462	3.5971
		(−2.6537***)	(−3.2237***)	(−1.9832**)	(−4.7779***)
	BVAR1	7.7311	10.0543	5.1025	13.4312
JOBU	DFM	2.3033	2.4067	2.4604	2.4403
		(−4.7008***)	(−1.7855*)	(−2.0133**)	(−4.2055***)
	BVAR3	10.0155	3.9097	7.3349	9.1191
KWAZ	DFM	3.3450	4.1419	4.0736	4.0927
		(1.0634)	(−1.8226*)	(−3.9921***)	(−3.2368***)
	SBVAR1	3.1487	6.2767	8.9245	8.0334
PRET	DFM	1.6371	1.7522	1.7688	1.7456
		(−1.6876*)	(−1.9747**)	(−1.9513**)	(−1.6536*)
	BVAR2	3.6546	4.4684	4.4186	3.3969
WCAP	DFM	2.0675	2.1097	2.2301	2.3201
		(2.4400**)	(2.4469**)	(1.9381**)	(2.4345**)
	BVAR1	0.4921	0.3304	0.9992	0.6025

Note: See note to Table I. In addition, BVAR2 and BVAR3 are the BVAR models based on the standard Minnesota prior with $w = 0.2$, $d = 1.0$ and $w = 0.3$, $d = 0.5$, respectively.

DOI: 10.1002/for

Table III. RMSEs for small middle-segment houses (2001:1 to 2006:4)

Region	Model	Quarters ahead			
		1	2	3	4
ECAP	DFM	4.7737 (–2.6052***)	5.1827 (–3.2060***)	4.8109 (–4.6989***)	4.7781 (–3.9490***)
	SBVAR1	9.5527	15.3944	19.7870	13.2875
JOBU	DFM	2.8092 (–1.9640**)	2.7369 (–2.7922***)	3.2401 (–2.3022**)	3.6976 (–3.1810***)
	SBVAR1	4.5326	7.3926	6.1887	12.9415
KWAZ	DFM	4.1300 (–1.7266*)	4.8269 (–3.5030***)	4.3750 (–1.6997*)	4.4162 (–3.0940***)
	BVAR1	5.9458	13.0952	5.4940	9.8042
PRET	DFM	2.4921 (–1.9651**)	2.9355 (–2.1687**)	2.6855 (1.6355)	2.5327 (–2.0016**)
	SBVAR1	4.3445	4.7499	1.7359	4.7650
WCAP	DFM	2.2799 (–1.6604*)	2.5541 (1.6566*)	2.7095 (1.2202)	2.6751 (–1.1130)
	SBVAR1	3.5421	1.7169	2.0921	2.9398

Note: See note to Table 1.

for Durban Unicity under the KwaZulu Natal metropolitan area, which in turn is forecast with the lowest errors by the BVAR model with $w = 0.1$, $d = 1.0$.

- *Medium middle-segment houses*. From Table II we learn that the DFM outperforms the best-performing model within the VAR category in four of the five metropolitan areas, with the exception of the first-quarter-ahead forecast for the Durban Unicity and all of the one- to four-quarters-ahead forecasts for Cape Town, for which the BVAR model with $w = 0.1$, $d = 1.0$ does the best. Unlike the case of large middle-segment housing, there does not exist a unique overwhelmingly favorite VAR model across all of the five metropolitan areas. The BVAR model with $w = 0.1$, $d = 1.0$ performed the best for the Eastern and Western Capes, while the BVARs with $w = 0.2$, $d = 1.0$ and $w = 0.3$, $d = 0.5$ stood out for Johannesburg and Pretoria, respectively. The SBVAR model based on the FOSC prior was the best-performing model for Durban Unicity.
- *Small middle-segment houses*. From Table III we observe that, as with the large and medium middle-segment housing, the DFM in general stands out as the best-suited model for forecasting house price inflation in all five metropolitan areas for all of the one- to four-quarters-ahead forecasts. The minor exceptions are the third-quarter-ahead forecast for Pretoria and the second- and third-quarter-ahead forecasts for Cape Town. Amongst the VARs, the SBVAR model based on the FOSC prior produces the lowest RMSEs for four of the five metropolitan areas, with the exception of KwaZulu Natal metropolitan area, for which the BVAR model with $w = 0.3$, $d = 0.5$ outperforms the other VAR models. Thus, as with the large middle-segment housing, the SBVAR model based on the FOSC prior is the overwhelming favorite within the category of VAR models.

Gupta and Das (2008) observed that the spatial models tended to outperform the other models for large middle-segment houses, while the unrestricted VAR and the BVAR models produced lower average out-of-sample forecast errors for middle and small middle-segment houses, respectively. In

DOI: 10.1002/for

our case though, in general, and especially for the large and small middle-segment houses, the SBVAR model based on the FOSC stands out once we take the DFM out of consideration.

When we take into account the cross-model tests of forecast accuracy proposed by Diebold and Mariano (1995), in the majority of cases where the DFM outperforms a specific type of the VAR model the statistics are significant at least at the 10% level. The exceptions are the third- and second-quarter-ahead forecasts for Pretoria and Western Cape, respectively, under the large middle-segment housing, and the fourth-quarter-ahead forecast for Cape Town for the small middle-segment houses. At the same time, in most cases where the alternative model tends to outperform the DFM, the Diebold–Mariano (1995)[4] test statistics are insignificant.

CONCLUSIONS

This paper analyzes whether the wealth of information contained in the DFM framework can be useful in forecasting regional house price inflation. As a case study illustration we use the DFM to predict house price inflation in five metropolitan areas of South Africa, namely Cape Town, Durban, Johannesburg, Port Elizabeth and Pretoria, using quarterly data over the period 1980:1 to 2006:4. The in-sample period contains data from 1980:1 to 2000:4, and the out-of-sample forecasts are based on one- to four-quarter-ahead forecasts over a 24-quarter forecasting horizon covering 2001:1 to 2006:4. The forecast performance of DFM is evaluated in terms of the RMSEs by comparing it with SBVAR models, based on the FOSC and the RWA priors, besides non-spatial models like the VAR and BVAR models with the Minnesota prior, estimated merely based on house price inflation of the five above-mentioned metropolitan areas of South Africa. Our results thus indicate that a data-rich DFM, in general, is best suited in forecasting regional house price inflation when compared to the alternative VARs.

APPENDIX: FOSC- AND RWA-BASED MINNESOTA PRIORS

FOSC prior

Given equation (5) in the text, i.e., $S_1(i, j, m) = [w \times g(m) \times F(i, j)]\frac{\hat{\sigma}_i}{\hat{\sigma}_j}$, and referring to the provincial map of South Africa given in Figure 1, the design of the F matrix based on the FOSC prior, discussed in the main text, given the alphabetical ordering[5] of the five metropolitan areas as the Eastern Cape Metropolitan area (Port Elizabeth/Uitenhage), Greater Johannesburg, the KwaZulu Natal Metropolitan area (Durban Unicity), Pretoria and the Western Cape Metropolitan area (Cape Town), can be formalized as in equation (A1). Note that each element of F represents the relationship between the respective pair of location, with 1.0 representing pairs that are neighbors, while

[4] If $\{e_t^{\mathrm{DFM}}\}_{t=1}^{T}$ denotes the forecast errors from the DFM model and $\{e_t^{\mathrm{ALT}}\}_{t=1}^{T}$ denotes the forecast errors from the alternative model, the Diebold and Mariano (1995) test statistic is then defined as: $s = \frac{1}{\sigma_l}$, where l is the sample mean of the 'loss differentials', $\{l_t\}_{t=1}^{T}$, using $l_t = (e_t^{\mathrm{DFM}})^2 - (e_t^{\mathrm{ALT}})^2$ for all $t = 1, 2, 3, \ldots, T$, and σ_l is the standard error of l. The s statistic is asymptotically distributed as a standard normal random variable and can be estimated under the null hypothesis of equal forecast accuracy, i.e., $l = 0$. Therefore, a negative value of s would suggest that the DFM model outperforms the alternative model in terms of out-of-sample forecasting.

[5] It must, however, be pointed out that alternative ordering of the five metropolitan areas do not affect our final results in any way.

DOI: 10.1002/for

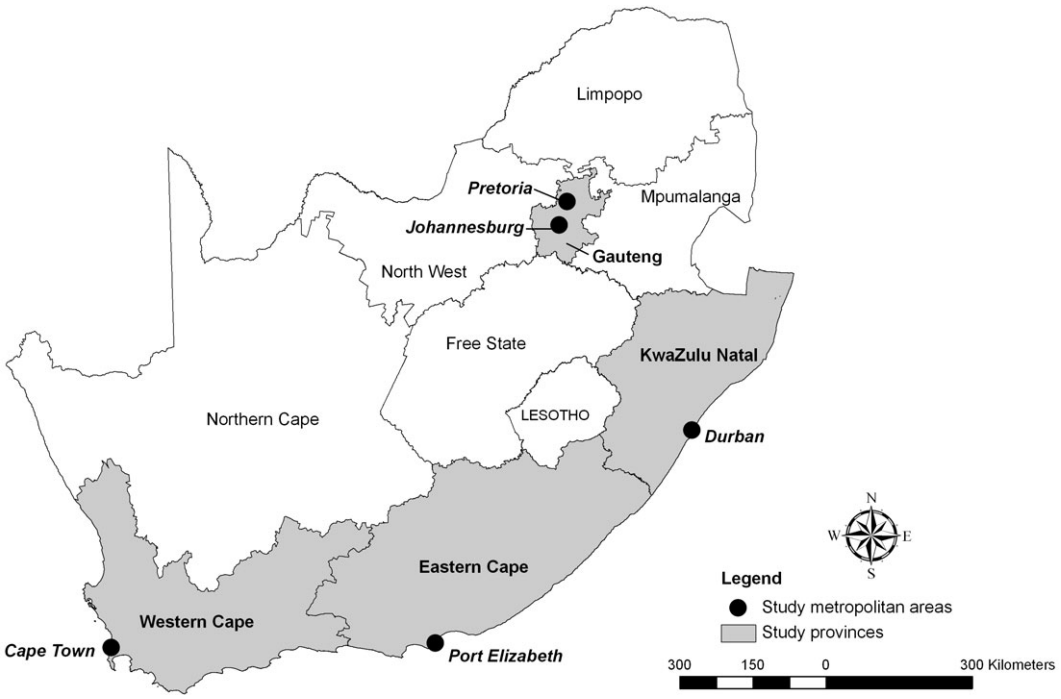

Figure 1. Provincial Map of South Africa. Source: CSIR, Pretoria, South Africa

0.1 represents non-neighbors. To illustrate the design of the F matrix, let us, for example, consider the first row of the matrix that corresponds to Port Elizabeth/Uitenhage in the Eastern Cape province. Eastern Cape has the provinces of KwaZulu Natal and Western Cape as immediate neighbors; hence the metropolitan areas of Durban Unicity and Cape Town that fall under these two provinces, respectively, have been treated as neighbors of Port Elizabeth/Uitenhage. Thus, speaking formally in terms of the entries in the different columns of the first row of the F matrix, we have a value of 1.0 on the third and fifth off-diagonal elements, besides the diagonal, since these two entries correspond to Durban Unicity and Cape Town, respectively. Since Johannesburg and Pretoria belong to the Gauteng province, which is not an immediate neighbor of the Eastern Cape province, the second and fourth columns of the first row of the F matrix have a value of 0.1. We follow a similar reasoning for the remaining entries of the F matrix, realizing that the immediate neighbor of Johannesburg is Pretoria and vice versa, while Port Elizabeth/Uitenhage is the only metropolitan area next to Durban Unicity and Cape Town. Mathematically, we have

$$F = \begin{bmatrix} 1.0 & 0.1 & 1.0 & 0.1 & 1.0 \\ 0.1 & 1.0 & 0.1 & 1.0 & 0.1 \\ 1.0 & 0.1 & 1.0 & 0.1 & 0.1 \\ 0.1 & 1.0 & 0.1 & 1.0 & 0.1 \\ 1.0 & 0.1 & 0.1 & 0.1 & 1.0 \end{bmatrix} \tag{A1}$$

DOI: 10.1002/for

RWA prior

To distinguish the F matrix under the RWA prior, discussed in the main text, from that in equation (A1), we denote it as F^1 as follows:

$$F^1 = \begin{bmatrix} 1.0 & 0 & 1.0 & 0 & 1.0 \\ 0 & 1.0 & 0 & 1.0 & 0 \\ 1.0 & 0 & 1.0 & 0 & 0 \\ 0 & 1.0 & 0 & 1.0 & 0 \\ 1.0 & 0 & 0 & 0 & 1.0 \end{bmatrix} \tag{A2}$$

The weight matrix given above is then standardized so that the rows sum to unity. Formally, we can write the standardized F^1 matrix, C, as follows:

$$C = \begin{bmatrix} 0.33 & 0 & 0.33 & 0 & 0.33 \\ 0 & 0.50 & 0 & 0.50 & 0 \\ 0.50 & 0 & 0.50 & 0 & 0 \\ 0 & 0.50 & 0 & 0.50 & 0 \\ 0.50 & 0 & 0 & 0 & 0.50 \end{bmatrix} \tag{A3}$$

Formally:

$$y_{it} = \delta_i + \sum_{j=1}^{n} C_{ij} y_{jt-1} + u_{it} \tag{A4}$$

On expanding equation (A4), we observe that multiplying y_{jt-1} containing the house price growth rates of five metropolitan areas at $t-1$ by the matrix C would produce a set of explanatory variables for each equation of the VAR equal to the mean of observations from the important variables (neighboring house prices) in each equation i at $t-1$. In other words, after normalization of the rows, one needs to calculate the neighborhood weighted average of the variable y_t from previous periods. This also suggests that the prior mean for the coefficients on the first own-lag of the important variables is equal to $1/c_i$, with c_i being the number of important variables in a specific equation i of the VAR model. However, as in the Minnesota prior, the RWA prior uses a prior mean of zero for the coefficients on all lags, except for the first own lags, and δ is estimated based on a diffuse prior.

Note that the RWA approach of specifying prior means requires the variables to be scaled to have similar magnitudes, as it does not make much intuitive sense to suggest that the value of a variable at t is equal to the average of values from the important variables at $t-1$. This transformation is not much of an issue as the data on the variables, in our case the house price inflation, can always be expressed as percentage change, or annualized growth rates, thus meeting the similar-magnitude requirements of the RWA prior.

As proposed by LeSage and Krivelyova (1999), a flexible form of the RWA prior standard deviations, $S_2(i,j,m)$, for a variable j in equation i at lag length m, is as follows:

DOI: 10.1002/for

$$
\begin{aligned}
S_2(i,j,m) &\sim N\left(\frac{1}{c_i},\sigma_c\right);\ j\in C;\ m=1;\ i,j=1,\ldots,n\\
S_2(i,j,m) &\sim N\left(0,\eta\frac{\sigma_c}{m}\right);\ j\in C;\ m=2,\ldots,p;\ i,j=1,\ldots,n\\
S_2(i,j,m) &\sim N\left(0,\rho\frac{\sigma_c}{m}\right);\ j\notin C;\ m=1,\ldots,p;\ i,j=1,\ldots,n
\end{aligned}
\tag{A5}
$$

where $0 < \sigma_c < 1$; $\eta > 1$ and $0 < \rho \leq 1$. For the variables $j = 1, \ldots, n$ in equation i, those variables that are important in explaining the movements in variable i ($j \in C$), the prior mean for the lag length of 1 is set to the average of the number of important variables in equation i, and to zero for the unimportant variables ($j \notin C$). With $0 < \sigma_c < 1$, the prior standard deviation for the first own-lag imposes a tight prior mean to reflect averaging over important variables. For important variables at lags greater than one, the variance decreases as m increases, but the restriction of $\eta > 1$ allows for the zero prior means on the coefficients of these variables to be imposed loosely. Finally, we use $\rho\sigma_c/m$ for lags on unimportant variables, which has prior means of zero, to indicate that the variance decreases as m increases. In addition, with $0 < \rho \leq 1$, we impose the zero means on the unimportant variables with more certainty.

ACKNOWLEDGEMENTS

Prof. Alain Kabundi gratefully acknowledges the financial support from Economic Research Southern Africa (ERSA), South Africa.

REFERENCES

Abraham JM, Hendershott PH. 1996. Bubbles in metropolitan housing markets. *Journal of Housing Research* **7**: 191–207.

Bai J, Ng S. 2002. Determining the number of factors in approximate factor models. *Econometrica* **70**: 191–221.

Bai J, Ng S. 2007. Determining the number of primitive shocks in factor models. *Journal of Business and Economic Statistics* **25**: 52–60.

Bernanke B, Gertler M. 1995. Inside the black box: the credit channel of monetary transmission. *Journal of Economic Perspectives* **9**: 27–48.

Boivin J, Ng S. 2005. Understanding and comparing factor-based forecasts. International Journal of Central Banking **1**: 117–151.

Burger P, Van Rensburg LJ. 2008. Metropolitan house prices in South Africa: do they converge? *South African Journal of Economics* **76**: 291–297.

Cho M. 1996. House price dynamics: a survey of theoretical and empirical issues. *Journal of Housing Research* **7**: 145–172.

Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–263.

Doan T, Litterman RB, Smis C. 1984. Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews* **3**: 1–144.

Dua P, Ray SC. 1995. A BVAR model for the Connecticut economy. *Journal of Forecasting* **14**: 167–180.

DOI: 10.1002/for

Elliott G, Rothenberg TJ, Stock J. 1996. Efficient tests for an autoregressive unit root. *Econometrica* **64**: 813–836.
Forni M, Hallin M, Lippi M. Reichlin L. 2000. The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics* **82**: 540–554.
Forni M, Hallin M, Lippi M, Reichlin L. 2003. Do financial variables help forecasting inflation and real activity in the euro area? *Journal of Monetary Economics* **50**: 1243–1255.
Forni M, Hallin M, Lippi M, Reichlin L. 2005. The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics* **82**: 540–554.
Gupta R, Das S. 2008. Spatial Bayesian methods of forecasting house prices in six metropolitan areas of South Africa. *South African Journal of Economics* **76**: 298–313.
Hamilton JD. 1994. *Time Series Analysis* (2nd edn). Princeton University Press: Princeton, NJ.
International Monetary Fund. 2000. *World Economic Outlook: Asset Prices and the Business Cycle*. IMF: Washington, DC.
Johnes G, Hyclak T. 1999. House prices and regional labor markets. *Annals of Regional Science* **33**: 33–49.
Kwiatowski D, Phillips PCB, Schmidt P, Shin Y. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? *Journal of Econometrics* **54**: 159–178.
LeSage JP. 1999. Applied econometrics using MATLAB, 1999. http://www.spatial-econometrics.com/html/mbook.pdf
LeSage JP, Krivelyova A. 1999. A spatial prior for Bayesian autoregressive models.*Journal of Regional Science* **39**: 297–317.
LeSage JP, Pan Z. 1995. Using spatial contiguity as Bayesian prior information in regional forecasting models. *International Regional Science Review* **18**: 33–53.
Litterman RB. 1981. *A Bayesian procedure for forecasting with vector autoregressions*. Working paper, Federal Reserve Bank of Minneapolis.
Litterman RB. 1986. Forecasting with Bayesian vector autoregressions: five years of experience. *Journal of Business and Statistics* **4**: 25–38.
Liu G, Gupta R, Schaling E. 2009a. Forecasting the South African economy: a hybrid-DSGE approach. *Journal of Economic Studies* (forthcoming).
Liu G, Gupta R, Schaling E. 2009b. A New Keynesian DSGE model for forecasting the South African economy. *Journal of Forecasting* **28**: 387–404.
Rapach DE, Strauss JK. 2007. Forecasting real housing price growth in the eighth district states. Federal Reserve Bank of St Louis. *Regional Economic Development* **3**: 33–42.
Rapach DE, Strauss JK. 2009. Differences in housing price forecast ability across U.S. states. *International Journal of Forecasting* **25**: 351–372.
Sims CA. 1980. Macroeconomics and reality. *Econometrica* **48**: 1–48.
Spencer DE. 1993. Developing a Bayesian vector autoregression model. *International Journal of Forecasting* **9**: 407–421.
Stock JH, Watson MW. 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* **97**: 147–162.
Stock JH, Watson MW. 2003. Forecasting output growth and inflation: the role of asset prices. *Journal of Economic Literature* **41**: 788–829.
Theil H. 1971. *Principles of Econometrics*. Wiley: New York.
Todd RM. 1984. Improving economic forecasting with Bayesian vector autoregression. *Quarterly Review, Federal Reserve Bank of Minneapolis* **Fall**: 18–29.
Topel RH, Rosen S. 1988. Housing investment in the United States. *Journal of Political Economy* **96**: 718–740.
Vargas-Silva C. 2008. The effect of monetary policy on housing: a factor augmented vector autoregression (FAVAR) approach. *Applied Economics Letters* **15**: 749–752.

Authors' biographies:
Sonali Das is a Senior Researcher (Statistics) at the Council of Scientific and Industrial Research, Pretoria. She has published referred papers in both theoretical and applied statistics in areas ranging from health, environment to housing.

DOI: 10.1002/for

Rangan Gupta is a Professor at the Department of Economics, University of Pretoria. He has published extensively in refereed journals, and his research interests are mainly macroeconomics and time series econometrics.

Alain Kabundi is an Associate Professor in the Department of Economics and Econometrics, University of Johannesburg. He has published extensively in refereed journals, and his research interests are mainly macroeconomics, financial economics and time series econometrics.

Authors' addresses:
Sonali Das, Logistics and Quantitative Methods, CSIR Built Environment, PO Box 395, Pretoria 0001, South Africa.

Rangan Gupta, Department of Economics, University of Pretoria, Pretoria, 0002, South Africa.

Alain Kabundi, Department of Economics, University of Johannesburg, Johannesburg 2006, South Africa.

DOI: 10.1002/for