FACTOR MODEL FORECASTS OF INFLATION IN CROATIA

Davor KUNOVAC[1]
Croatian National Bank, Zagreb

Article*
UDC 336.748.12(497.5)
JEL E44

Abstract

This paper tests whether information derived from 144 economic variables (represented by only a few constructed factors) can be used for the forecasting of consumer prices in Croatia. The results obtained show that the use of one factor enhances the precision of the benchmark model's ability to forecast inflation. The methodology used is sufficiently general to be able to be applied directly for the forecasting of other economic variables.

Key words: factor models, time series analysis, inflation, forecasting

1 Introduction

1.1 Motivation

In this paper we test whether some elements of factor analysis and a related mathematical method - *Principal Component Analysis (PCA),* can be used in the forecasting and better understanding of economic processes in Croatia.

The idea that the correlation structure of a large number (hundreds or thousands) of variables is accurately approximated by using only a handful (often only five or six) of the variables, or factors, is found in research conducted in the area of both natural and social disciplines. Roots of factor analysis can be found in psychology, where, at the beginning of the last century, Charles Spearman (Spearman, 1904) conducted the extraction of fac-

* *Received:* June, 1 2007.
Accepted: December, 10 2007.

[1] The views expressed in this paper are those of the author and do not necessarily reflect the views of the Croatian National Bank. The author hereby wishes to thank anonymous referees for their useful comments

tors in the context of the measurement of intelligence on the basis of examination results in mathematics, foreign languages, and so on. As well as by psychologists, factor analysis is also often used by chemists – chemometricians and experimental physicists.

Standard econometric models, such as VAR models or simultaneous equation systems, due to the short time spans of available data sets, can be used for simultaneously modelling the interaction of only a handful of variables - usually fewer than ten. Therefore, in the last few years intensive research has been conducted on the possibility of *the compression* of economic and financial data. In the context of forecasting key macroeconomic variables, US economists James H. Stock and Mark W. Watson (Stock and Watson 1998; 1999; 2002) in a series of papers have investigated the exploitability of factor models. Given a large number of macroeconomic variables, they extracted a few factors which (in a measurable quantity) summarise information of the entire US economy. In the second step the factors are used for forecasting industrial production, inflation, and so on. The accuracy of their factor forecasts is at least comparable with the forecasts of standard econometric models - univariate regressions, autoregressions, VARs and models based on leading indicators. This triggered further testing of the usefulness of factor models (i.e. Matheson, 2006, Camacho and Sancho 2003, Camba - Méndez and Kapetanios, 2005). These results are promising but should not be automatically applied to other economies without prior direct empirical evaluation and possible adjustments. The uncritical appropriation of these results by young, open economies would be deemed especially misleading, firstly due to the differences in quality and quantity of available statistics and then to the nature of the functioning of the mechanisms of particular economies. Therefore, this analysis is conducted in order to examine the usefulness of a relatively new technique in modelling and forecasting economic processes in Croatia. Furthermore, the contribution which such an analysis can offer in the context of a general overview of forecasting performances of factor models, so far not standard, is also clear.

1.2 Applications of factor models in macroeconomic analysis

During the last decade, there has been a continuous increase in the number of papers dealing with factor models in economics. A brief overview of recent papers determining the key directions of the development of economic factor models from the aspects of both methodology and empirical analysis is given below.

In addition to the work of Stock and Watson, two more papers should be mentioned in the context of methodology development. Forni et al. (2000) provides a method for factor forecasting based on elements of spectral analysis and as such, it is considered to be complementary to the standard Stock-Watson methodology. Furthermore, Bai-Ng (2002) define the estimator of the number of unknown factors used in the forecasting model, which is considered to be one of the basic decisions when designing factor models.

Moreover, Bernanke and Boivin (2003) test results of factor analysis in the context of needs of the US Central Bank and give an interesting critique of empirical analysis of monetary policy, which usually presumes that decisions of central bankers are grounded on only a few macro variables, such as inflation or GDP, while in practice the number of variables analysed is significantly larger. Along the same line, Bernanke and Boivin estimate the reaction function of FED, determined by overall information of a large number of relevant economic variables. This overall information is represented by only a few estimated factors, which are interpreted as the forces generating the entire economy.

Factor models have shown to be very useful in a *flash estimate* of GDP. It is known that official data for GDP are available only with a few months delay, but it is of interest to know the current state of economic activity, which thus has to be estimated. In this context, an important contribution has been made by the analysis conducted by Schumacher and Breitung (2006) and Giannone et al. (2005).

Economic applications of factor models are mostly non theoretical in nature – they are related to forecasting. One of the exceptions is a paper by Boivin and Giannoni (2005) where information contained in a large number of variables is used for parameters estimation of theoretical DSGE[2] models.

Analyses testing credibility of factors models constructed from relatively large number of series but given a short time span are of special importance for this paper. In this context we can refer to two analyses. Banerjee et al. (2006) examine the stability of factor forecasts in a small sample environment with present structural breaks, where Slovenia is considered a representative of the new EU member states. Furthermore, J. Boivin and S. Ng (2005) using Monte Carlo simulations, question the sensitivity of the quality of factor forecast, given the length of time series available. This analysis encourages the implementation of factor analysis even in the conditions of short time series such as those obtaining in Croatia.

The EC analysis (Grenouilleau, 2006) and the economic indicator of the Chicago Fed (Chicago Fed Letter, Number 151) could also be distinguished as important examples of the implementation of the factor model in economics.

Finally, an overview of the results of factor forecasts of inflation and real activity contained in a number of relevant papers (46) is given in a meta-analysis by Eickmeier and Ziegler (2006).

1.3 Analysis structure

The structure of this paper is as follows. The next section contains the main mathematical results necessary for our factor model; firstly, the elements of the Principal Component Analysis and the construction of the principal component of variables from the group of prices and exchange rates will be laid down.

In Section 3, *Stock-Watson* forecasting methodology will be applied to inflation in Croatia, where factors and parameters will be estimated via principal components. Furthermore, possibilities for future upgrading of the model will be elaborated as well.

A conclusion is given in Section 4.

[2] Dynamic Stochastic General Equilibrium.

2 Principal components

2.1 Definition

Let us assume that there are given observations for N variables $X_1 K, X_N$, the covariance structure of which is of interest.[3] In our context, at any given moment, the economy is approximated with the N - dimensional random vector.

In order fully to describe the covariance-variance structure of these variables, data on all N variables are needed. However, it is often possible to describe a large part of data variability by k $(< N)$ components - the principal components. Let us formalise this.

Let $X^\tau = (X_1, ..., X_N)$ be a random vector with covariance matrix $\sum$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N \geq 0$ For arbitrary real vectors a_i, $i = 1, ..., N$ we can define linear combinations:

$$Y_i = a_i^\tau X = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{iN}X_N, \quad i = 1, \ldots, N. \tag{1}$$

These linear combinations of variables $X_1, ..., X_N$ are random variables with variance and covariance:

$$VarY_i = a_i^\tau \Sigma a_i, \quad i = 1, \ldots, N \tag{2}$$

$$Cov(Y_i, Y_k) = a_i^\tau \Sigma a_k, \quad i, k = 1, \ldots, N. \tag{3}$$

The principle components are defined recursively as uncorrelated linear combinations of $Y_1, Y_2, ..., Y_N$ (1) having the maximal variance. In order to ensure uniqueness, in the definition of principal components we additionally need the weight vectors a_i to have a unit norm, therefore $a_i^\tau a_i = 1$:

- the first principal component $PC1$ is linear combination $a_1^\tau X$ with maximal variance, given condition $a_1^\tau a_1 = 1$,
- the second principal component $PC2$ is linear combination $a_2^\tau X$ with maximal variance, given conditions $a_2^\tau a_2 = 1$ and $Cov(a_1^t X, a_2^\tau X) = 0$,

$\vdots$

- i-th principal component (for $i \leq N$) PCi is linear combination $a_1^\tau X$ with maximal variance given conditions $a_i^\tau a_i = 1$ and $Cov(a_i^\tau X, a_k^\tau X) = 0$, for all $k < i$.

By the set definition, using available data we generate non-correlated components of maximal variance. From a practical point of view, two questions arise. First, how can we

[3] In factor analysis we examine variances and covariances of data, therefore only the second moment information. Possible useful information of third or higher moments is ignored.

efficiently construct the principal components, and second; which share in the total variance, that is, the share in the number $\sum_{i=1}^{N} VarX_i$, is described by the first *k* principal components. From this aspect we provide two central results, however, without proof (available in Johnson and Wichern, 1998).

Result 1. Let $\sum$ be the covariance matrix of vector $X^\tau = (X_1, ..., X_N)$. Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N \geq 0$ be eigenvalues and, *a* $x_1, ..., x_N$ the corresponding (unit norm) eigenvectors of $\sum$. Then the i-th principal component is given by:

$$Y_i = x_i^\tau X, \quad i = 1, \dots, N. \tag{4}$$

Furthermore:

$$VarY_i = \lambda_i, \quad i = 1, \dots, N \tag{5}$$

$$Cov(Y_i, Y_k) = 0, \quad i \neq k. \tag{6}$$

Result 2. Let $Y_1 = x_1^\tau X, \dots, Y_N = x_N^\tau X$ be the principal components (from the result 1). Then, the following holds:

$$\sum_{i=1}^{N} VarX_i = \lambda_1 + \dots + \lambda_N = \sum_{i=1}^{N} VarY_i. \tag{7}$$

The first result says that the principal components are calculated as a product of the vector *X*, and the corresponding covariance matrix eigenvectors and the second result ensures equality of the sum of the variances of the elements of the original vector *X* and of constructed principal components. Furthermore, the share of the *k*-th principal component's variance in the total variance is equal to:

$$\frac{\lambda_k}{\lambda_1 + \dots + \lambda_N}, \quad k = 1, \dots, N \tag{8}$$

Consequently, the share of the first *l* components is:

$$\frac{\lambda_1 + \dots + \lambda_l}{\lambda_1 + \dots + \lambda_N}, \quad l \leq N. \tag{9}$$

From the operational point of view, it is sufficient to calculate the sample covariance data matrix and its eigenvectors and eigenvalues. Usually, when calculating eigenvalues of high dimensional matrix, appropriate program support is deemed necessary. From the numerical stability point of view, instead of direct calculation of eigenvalues and eigenvectors, it is better to conduct *Singular Value Decomposition*. Details of this process can be found in some of the textbooks on (numerical) linear algebra. PCA is more useful as the expression (9) is closer to the unit for relatively small *l*, that is, if we can *capture* most of variability of the data using only a few components. Which *l* is small enough, and how close to unity is indeed close enough, depends on the nature of the problem and the discretionary decision of the individual, although there are certain formal procedures which help in solving this question.

In application, we usually have T realisation of vector X observed as matrix X ($N \times T$). Let Y ($K \times T$) be the matrix of the first K principal components. If for some K a significant share of the total data variance is explained by the first K components ($K<N$), then the N-dimensional vector of the original data $X^{\tau} = (X_1, ..., X_N)$ may be represented by K-dimensional vector $Y^{\tau} = (Y_1, ..., Y_N)$. Such compressed data can be used in further statistical analysis. Therefore, with the reduction of the dimension we ignore a certain quantity of information in order to be able to use information of a potentially large number of time series in modelling. Effectiveness of this *trade-off* is questioned from case to case. This is particularly useful in conditions of short time series where, for example, one can use regression analysis to measure the influence of a large number of variables (reduced to a small number of constructed components) on key series. In this context, it is worth examining the potential of PCA in economic research in Croatia.

2.2 First principal component of prices and exchange rates

With the aim of illustration, we will analyse the first principal component of the set of 32 prices and exchange rates series we consider to be relevant for the dynamics of Croatian Consumer Price Index (CPI). The appendix includes a complete list of all series.

We will use monthly data, starting from January 1998 to September 2007.[4] Prior to the calculation of his covariance matrix, data are transformed as follows:

1 Program X12 - ARIMA is used to filter the data[5] and in further analysis only trend-cycle components[6] are taken into consideration.

2 In the second step we consider log differences of the data.[7] These differences approximate monthly growth rates.

3 We standardise data in such a way that all series have zero expectation and a unit standard deviation. In this way, regardless of the measurement unit, we are putting them on the same scale. Notice that covariance matrix of standardised data is the correlation matrix, and therefore the principal components analysis of standardised data is in fact an analysis of the correlation structure of data.

Following this transformation, data are prepared for classical principal components analysis. Firstly, we calculate the sample covariance (correlation) matrix and its eigenvalues and eigenvectors. Using these mathematical objects, following the previously de-

[4] The Croatian Statistics Bureau started publishing CPI in February 2004 and the index starting from January 1998 was calculated afterwards.

[5] This is an official SA program used by the US statistical office. Details on this method can be found on the web site: http://www.census.gov/srd/www/x12a/

[6] Additive decomposition of the series $Y = TC + S + I$, is assumed, where TC is trend-cycle component (T - trend, possibly cycle of markedly low frequency, C is cycle, that is, medium term component), S is seasonal component and I is the residual, that is, irregular component. The use of trend-cycle component, therefore disregard of both the seasonal and irregular component is motivated by discussion in Camacho and Sancho (2003) in the context of application of factor models on Spanish data.

[7] Natural logarithm is used in this case

scribed procedure, we calculate principal components and comment on the structure of the first component with the aim of further insight of the correlation level between the variables in the group. Table 1 gives the basic results for the first few components.

Table 1 First six principal components of prices and exchange rates series (PCi, i = 1,...,6, denotes i –th component)

Component	PC1	PC2	PC3	PC4	PC5	PC6
eigenvalue	8.49	5.46	4.07	2.62	1.82	1.58
variance proportion	0.28	0.18	0.14	0.09	0.06	0.05
cumulative variance proportion	0.28	0.47	0.60	0.69	0.75	0.80

Source: authors' calculations

The first six components explain 80% of the total data variance while the first ten components describe approximately 92% of the variance. By construction, the first component captures the major share of the variance - in our case 28%. Due to better insight, we would point out that the 10th component describes only 2% of data variance. Given the calculated values it can be concluded that there is a relatively strong correlation between growth rates within the group of series. By definition, the first principal component is a maximum variance linear combination of all series of the observed group and therefore it could be considered an index constructed from 32 prices and exchange rates time series. For the purpose of interpretation, corresponding weights of this index that is, eigenvectors of correlation matrix, should be analysed. Let us notice that the sign of individual ponders in that context does not help, due to the fact that principal components are by definition identified only up to the sign. Following this, we should concentrate only on weights' absolute values.

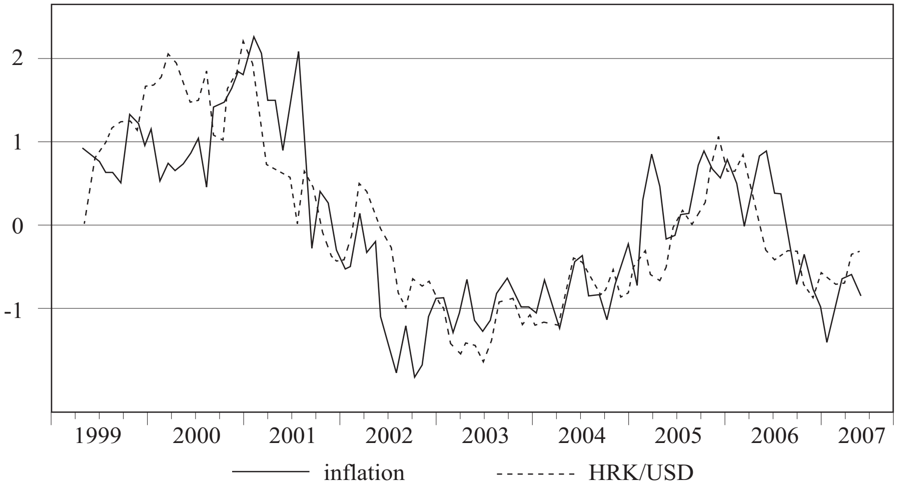
Table 2 Weights in the first component and components correlation with four most important series

	IPC	Goods (excl. energy)	Goods	HRK/USD
weight	0.31	0.29	0.29	0.25
correlation with component	0.85	0.80	0.80	0.70

Source: authors' calculations

Table 2 contains data for four series from the observed group of prices and exchange rates which entered into constructed index (first principal component) with the highest weights. In addition, correlation of these series with the calculated index is shown. Three series with highest weights are price indexes - *total index of consumer prices* (0.31), GOODS, *excl. energy* series (0.29) and *GOODS* (0.29). These results were expected since a large part of observed group of variables consists of CPI sub indexes, which are, as such, correlated to CPI. By definition, out of all possible linear combinations of the given variables of prices and exchange rates, their first principal component optimally approximates the total variance of the group. If there is a strong correlation between its elements, it is to be expected that the principal component also follows this common dynamics. Interestingly, the fourth variable[8] is the average monthly exchange rate *HRK/USD*. For the purpose of illustration of the correlation of movement of consumer prices and kuna exchange rate to US dollar, figure 1. shows standardised yoy growth rates of average monthly exchange rate of *HRK/USD* and the consumer price index. More details on relation of kuna exchange rate and US dollar can be found in Jankov et al. (2007).

Figure 1 Standardised yoy growth rates of average monthly HRK/USD exchange rate and CPI



Source: authors' calculations; HNB

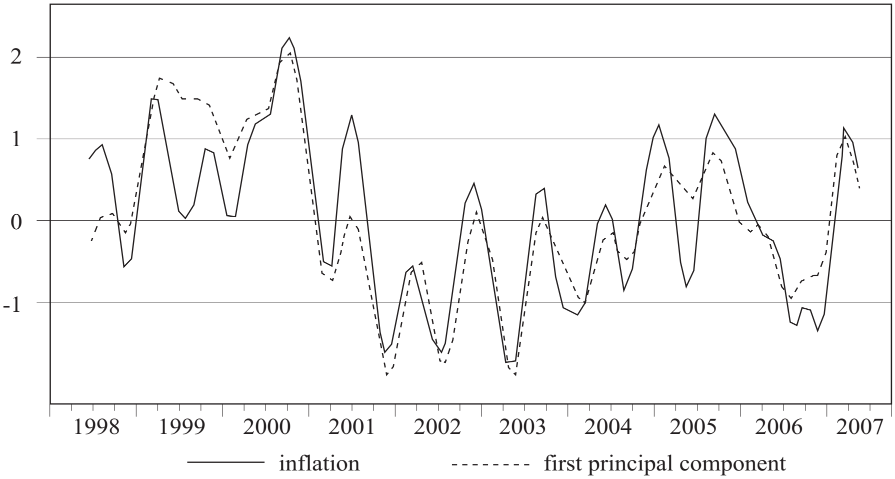
Figure 2 shows the constructed first principal component of prices and exchange rates series together with monthly growth rate of the trend-cycle component of the CPI.

The principal component is by construction a linear combination of monthly growth rates of trend-cycle series of the observed group. Therefore, this index, which encompasses, in addition to CPI *components*, a variety of other variables relevant for inflation movement, explicitly and measurably takes into account their dynamics.

This example illustrates how the methods of compression of the information can be used for index design, according to clear criteria that it should represent the optimal one-dimensional representation of a group of variables. Indexes of this type are used for construction of indicators of core inflation, level of economic activity and so on.

[8] We can notice (table 2) that the same ordering as in the case of weights remains if we look at correlations, but this is not always the case (see, for example, Johnson and Wichern, 1998).

Figure2 First principal component and standardized monthly CPI inflation



Source: authors' calculations; HNB

3 Factor forecast of inflation

3.1 Introduction

In the introduction, we mentioned that in the course of forecasting the key variable in central banks, a potentially huge number of economic variables are taken into consideration. The central issue in this context is how to efficiently compress the information contained in a large number of variables. A formal framework for such compression of information and its further application in statistical analysis has been proposed by, among others, Stock and Watson (1998; 1999; 2002).

In this section, we firstly define a factor model and describe how factor analysis in general can be used in the analysis of a large number of more or less related variables. After that, the model is applied to Croatian data in an inflation forecasting exercise.

3.2 Definition and basic features

In the previous section, we put forward the basic facts of principal components analysis. Factor analysis is based on a different idea but under certain assumptions, the parameters can be estimated using principal components analysis.

Let us assume that $T \times N$ matrix X, contains the realisations of N variables $X_1, ..., X_N$ during T time periods. Factor analysis is based on the assumption that there are r ($<N$) variables - *factors*, $f_1, ..., f_r$, which are sufficient for modelling given N variables. Now, let us assume linear dependence of original data contained in X on factors where matrix Λ contains coefficients. Therefore, let us assume the following:

$$X = F\Lambda^{\tau} + \varepsilon, \tag{10}$$

where $X\,(T\times N)$ is standardised matrix of original data, F is $(T\times r)$ matrix of factors, Λ $N\times r$ is *factor loadings* matrix and ε is the noise matrix. By elements we have:

$$X_{ti} = F_t\lambda_i^{\tau} + \varepsilon_{ti} \quad t = 1,\ldots,T,\ i = 1,\ldots,N, \tag{11}$$

where F_t is r-dimensional vector of factors at time t, and λ_i is vector of coefficients which define dependence of the i-th variable on factors. This is a so-called static representation of factor model where dependence of data on only contemporaneous factors is assumed.[9]

We see that the factor model assumes that each variable can be decomposed on linear combination of factors and the error term. In this context, factors represent driving forces governing the whole economy and therefore being common for all observed variables in X. Specificities of individual variables are contained in error terms, containing both *true error* and possibly factors specific for that particular variable.

In (11) a certain correlation between the components of $\varepsilon$is allowed. Therefore, presented model is also called approximate factor model (Bai-Ng 2002).

In equations (10) and (11) all but data matrix X. are unknown. Therefore, in order to estimate matrix Λ ($N\times r$ parameters) and factors ($T\times r$ parameters), it is necessary to impose certain restrictions on parameters. Therefore let us assume[10] that $\Lambda^{\tau}\Lambda = I$.

Parameters are estimated applying the least squares, where we look for the estimates of $\hat{F}$ andi $\hat{\Lambda}$ which minimise the error (11) that is, for which $\sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{ti}^2$ is minimal, under condition $\Lambda^{\tau}\Lambda = I$. This is a classic problem of principal components[11] which is solved by choosing $\hat{\Lambda}$ as eigenvectors corresponding to first r maximal eigenvalues of matrix $X^{\tau}X$. Then it holds that $\hat{F} = X\hat{\Lambda}$. The estimation details are given in the appendix, while more information on the properties of this estimator can be found in Stock-Watson (1998).

The procedure of factor estimation using principal components method is straightforward; however a few issues arise in the course of the implementation. First of all, how should one select a number of factors, that is how can one estimate r (Bai-Ng, 2002)? The consistence of data with the assumed structure (10) is also not clear a priori. Furthermore, there is a question of identification in (10), which is closely related to factor interpretation. That is, for $r\times r,$ matrix O for which $OO^{\tau} = I$, (10) can be also rewritted as:

$$X = (FO)(O^{\tau}\Lambda^{\tau}) + \varepsilon, \tag{12}$$

[9] Beside static factor models there are also dynamic factor models where it is assumed that elements from X in addition to present factor values also depend on their lagged values. However, dynamic models have static representation of the type (10) (Stock and Watson 2002), therefore, here we consider static models only.

[10] Alternatively, we can assume orthogonality of the factor – $F^{\tau}F = I$ which leads to the same solution (appendix).

[11] Besides the principal components method, parameters of factor models can be estimated from state *space* representation using *maximum likelihood*. However, maximisation becomes very unreliable as N (the number of variables out of which we are extracting factors) grows. Therefore, it is used only when extracting factors given a smaller number of variables.

Now, *new* factors are given by FO and factor *loadings* by $O^{\tau}\Lambda^{\tau}$. However, this is not a problem if factor analysis is used only in the preparation of data for a forecasting model because each pair of FO and $O^{\tau}\Lambda^{\tau}$ generates the same X. The problem occurs if we wish to attribute structural interpretation to the factors because it is not a priori clear whether F represents true factors (forces generating the entire economy) or whether it is just a transformation, which is very hard to interpret.

3.3 Forecasting model

Here, we apply the *Stock-Watson* forecasting model to Croatian economic series. The aim is, to forecast the value of variable y at $T + h$ using data available at time T, that is, to forecast y_{T+h}, for a certain h. In our case, we are trying to answer the question of whether the information provided from 144 quarterly series today (see appendix), can help in forecasting inflation h steps ahead?

The model is following:

$$X_{ti} = F_t \lambda_i^{\tau} + \varepsilon_{ti} \tag{13}$$

$$y_{t+h} = \alpha_h + \beta(L)F_t + \gamma(L)y_t + \varepsilon_{t+h}, \tag{14}$$

where α_h is a constant, $\beta(L)$ and $\gamma(L)$ are finite-order polynomials in variable L, where L is a standard lag (backward shift) operator. X_{ti} is the value of i-th variable at time t and F_t is r-dimensional factor vector at t. Therefore in (14) we assume that there is a linear dependence between scalar variable *y at time t + h* and factors *up to t*. Additionally, we assume that something on the future of the variable of interest can be said using information form its own history, therefore, the term $\gamma(L)\gamma_t$ also included.

We cannot start with the estimation of the parameters in (14), because not all the predictors (factors F_t) are known. We cannot observe true values of factors, but we estimate them from (13) according to the described procedure. Therefore we adopt the following strategy:

1 Factor estimation. From (13) we estimate time series of factors.

2 Forecasting. Up to the moment T, data are known. Regression coefficients $\hat{\alpha}_h$, $\hat{\beta}(L)$ and $\hat{\gamma}(L)$ are estimated using data from time 1 to T (using least squares). Now, forecasts are constructed as $\hat{y}_{t+h} = \hat{\alpha}_h + \hat{\beta}(L)\hat{F}_t + \hat{\gamma}(L)y_t$. In order for this forecast to be the best in the MSE sense, we assume that $E(\varepsilon_{t+h} \mid I_t) = 0$, where I_t is information available up to the time t.

The *forecasting* step needs some explanation. Forecasting is conducted with direct projection from data available up to the moment t, therefore, as $\hat{y}_{t+h} = \hat{\alpha}_h + \hat{\beta}(L)\hat{F}_t + \hat{\gamma}(L)y_t$. This is suitable because now it is not necessary to forecast factors (for example, using VARs) in order to find $\hat{y}_{t+h}$. It is not a priori clear which method yields better results, but the procedure conducted here is simpler and, more importantly, requires the estimation of a smaller number of parameters.

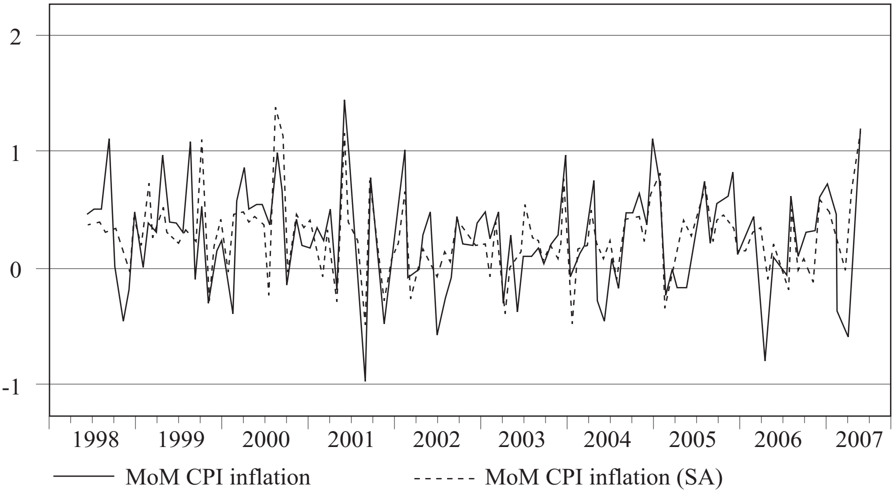
Data set under consideration is divided in five groups (sectors) – *Prices and exchange rates* (32 series), *Labour market* (24 series), *External sector* (28 series), *Real sector (*26 series) and *Money* (34 series) which amounts to 144 series. For a list of all series see the appendix.

Analogously as in the calculation of principal components series are first transformed. The theory requires that the series which enter the analysis are stationary. Under the assumption that the original data are integrated of order one (therefore stationary after differenced only once) with the conducted transformation, stationarity is approximated.[12] Seasonally adjustment is used to *clean up* the seasonal component, logarithms reduce possible heteroskedasticity, and differencing eliminates trends.

Forecasting the CPI inflation is of our primary interest. This index has a prominent seasonal component, therefore, prior to the analysis, this series should be seasonally adjusted. Figure 3 shows monthly (mom) and annual rates (yoy) of original and SA CPI.

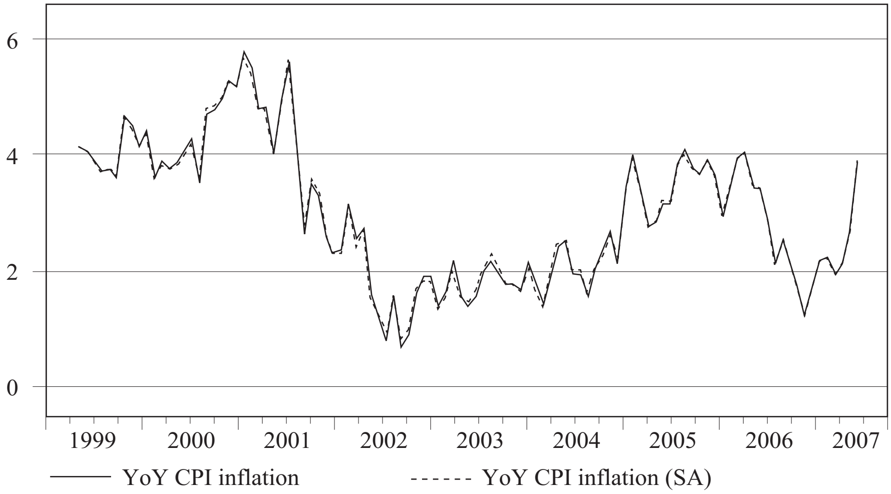
We see that monthly rates strongly depend on whether the series is seasonally adjusted, while in the case of annual rates this is less important as the seasonal component is closer to some periodical function.[13]

Figure 3 Growth rates for original series and seasonally adjusted series of consumer prices index. The first figure shows monthly and the second annual rates



[12] With the conducted transformations we really only approximate stationarity. Differencing is the most questionable and it is very hard to determine how many times a single series needs to be differenced. All series which are included in calculation (except exchange rates) exhibit clear rising or declining trends which suggests that they should be differentiated at least once. Possible need for further differencing is hard to determine in a short time series environment, having in mind the low power of *unit - root* tests.

[13] Let us assume that the logarithm of a certain variable, recorded on a monthly frequency is decomposed on seasonal component and (*S*) the leftover part (*N*): $lnY_t = S_t + N_t$. Annual growth rates are approximated with $lnY_t - lnY_{t-12} = S_t + N_t - (S_{t-12} + N_{t-12}) = S_t - S_{t-12} + (N_t - N_{t-12})$, and, if *S* is a periodical function of period 12, that is $S_t = S_{t-12}$, annual growth rates do not depend on seasonal component, so it is not important whether the series is seasonally adjusted or not (figure 3).



Source: authors' calculations; HNB

With the data available up to the time *t*, we are interested in finding the expected growth rate for variable *z*, at *t* + *h*, therefore, the expected value of

$$y_{t+h}^{h} = \frac{400}{h} \ln \frac{z_{t+h}}{z_t}. \tag{15}$$

Rates defined in this way are annualised, and forecasts for various horizons *h* are directly comparable. Let us additionally define (annualised) quarterly rates for series of our interest $y_t = 400 \ln \frac{z_t}{z_{t-1}}$. Let us notice that (15) gives accumulated quarterly rates for the period from *t* + 1 to *t* + *h*.

Our final goal is factor forecasts of the *consumer price index*. In order to evaluate their quality as accurately as possible, it is necessary to simulate real conditions of forecasting to the highest possible degree.

We use quarterly data, from the interval from the first quarter of 1998 to the second quarter of 2007. For *h* steps ahead forecast, the interval from the first quarter of 1998 to (4 - *h* + 1)-th quarter 2003 is used only for estimation while the rest of the data is also used for forecasting in the following recursive process:

1 Principal components methods is used to estimate factors (i.e. to calculate $\hat{F}_t$).

2 Akaike's information criterion is used to determine lag length both for factors and dependant variable

3 Least squares are used to estimate parameters of regression $y_{t+h}^{h} = \alpha_h + \beta(L)F_t + \gamma(L)y_t + \varepsilon_{t+h}$.

4 Forecast is constructed as $\hat{y}_{t+h}^{h} = \hat{\alpha}_h + \hat{\beta}(L)\hat{F}_t + \hat{\gamma}(L)y_t$.

5 We add a new observation to the sample and go to 1 (except when the sample is exhausted; in that case the procedure is over).

In this manner, forecasts for arbitrary *h* are calculated for the period from first quarter 2004 to second quarter 2007. The algorithm leaves the possibility of defining several types of the model. Using our data we test the quality of the following two models:

Autoregressive model – AR. Let us assume that rates y_{t+h}^{h} depend exclusively on quarterly rates (and the lagged values) $y_t = 400 \ln \frac{z_t}{z_{t-1}}$. The lag length, from 0 to 4, is selected by AIC. This model is used as *benchmark*. Discussion on the quality of forecasts given by this model can be found in Kapetanios et al. (2007).

Autoregressive factor model – ARF. Let us assume that y_{t+h}^{h} depend on both current and lagged values of factors and rates y_t. The lag length is selected by AIC (from 0 to 4 for *y* and from 0 to 3 for factors). Let us define the model for *k* factors, *k* = 1,2, therefore two ARF models.

3.4 Forecasts evaluation

The quality of forecasts is measured by the standard measure - Mean Squared Error (MSE) defined in the following way. Let $x_1, ..., x_N$ be actual (observed) values, and $\hat{x}_1, \ldots, \hat{x}_n$ forecasts of random variable *x*. Then:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \tag{16}$$

In table 3 we give MSE statistics for inflation forecasts for horizons from 1 to 4. Statistics are given as ratios to MSE of the autoregressive (AR) model. Therefore, when the value is higher than unity, the model (in MSE - context) is worse than benchmark model. However, the benchmark defined here is not trivially defined (for example, as a random walk model etc) and cannot be beaten easily. Results suggest that factors extracted from observed group of series have a certain potential in forecasting the dynamics of consumer prices index. Adding a single factor (model ARF1) to a certain extent enhances forecasts in relation to the benchmark. However, further adding factors did not improve forecasts.[14] This result is in line with the results from Stock-Watson (2002) or Matheson (2006).

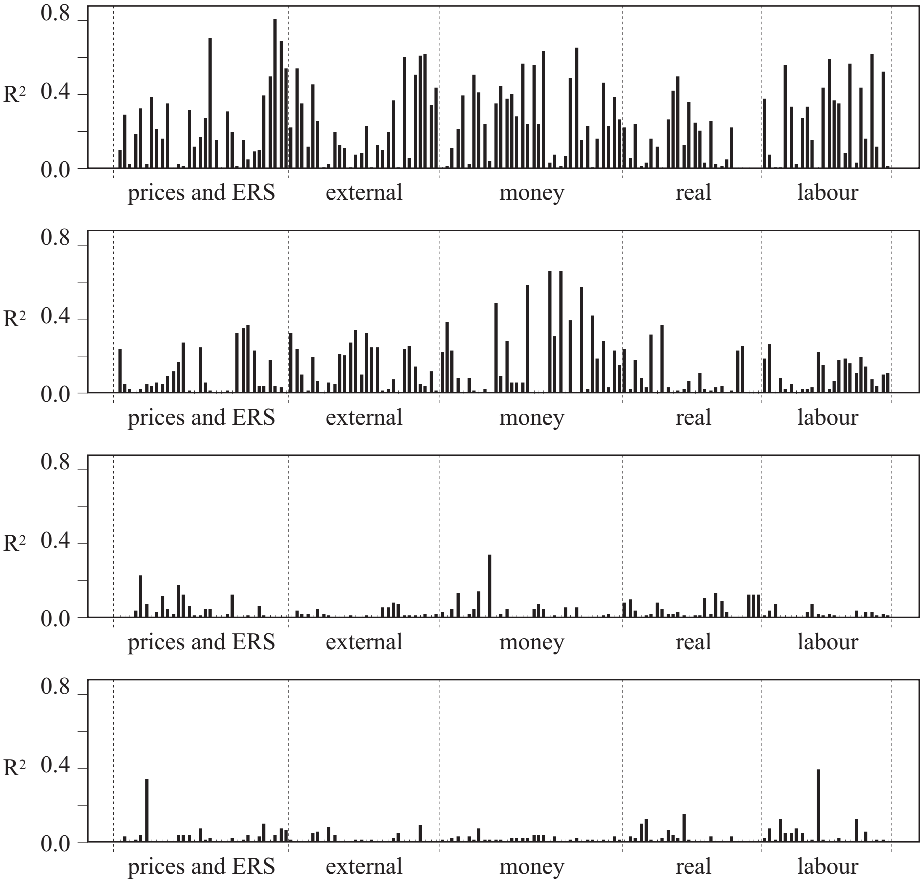
Table 3 Results of inflation forecasts of consumer prices index. MSE statistics are shown as ratio to benchmark

Horizont	1	2	3	4
ARF1	0.91	0.81	0.72	0.92
ARF2	0.99	1.05	3.55	2.87

Source: authors' calculations

[14] Analysis is also conducted on data on a monthly frequency where forecasts are constructed on horizons up to 12 months ahead. Results of this analysis suggest that factor forecasts with the use of two factors are better than benchmarks on short horizons – up to 7 or 8 months. Although longer series of data are available with monthly frequency, the aim of the analysis is the construction of inflation forecasts up to a year ahead. With the given scope of available data, quarterly forecasts for up to four periods ahead seem to be more credible than monthly forecasts for up to twelve periods ahead.

Figure 4 Coefficients of determination from regressions between each of 144 series on first, second, tenth and eleventh factor



Note: Vertical lines separate series by sectors. Statistics (height of columns) for every factor show the extent of variance for particular series which the particular factor explains. First two factors cumulatively explain approximately 45% of the overall variance of all series. For illustration, the tenth and eleventh factors jointly explain only 4% of the variance.

Source: authors' calculations

3.5 Further improvement of the model

Using a single factor increases the accuracy of the inflation forecast compared to the benchmark. However, there are many possibilities for improvement of the factor model implemented. We will outline only some of the segments of analysis that were partially or entirely ignored. Therefore, with this outline we give recommendations for future work.

- *Determining the number of factors to be included in the forecast model.* Number of factors should be determined using some of the information criteria developed for this purpose. In this respect the most reliable is the Bai-Ng estimator (Bai-Ng, 2002). In this analysis we have tested models in which we have included up to two factors. Cumulatively, they account for 45% of the total sample variance. Figure 4 illustrates a relatively strong correlation of the first two factors and all series and a very low correlation of the tenth and eleventh with the series. This partly justifies the *ad hoc* method of selection of the number of factors.
- *Treatment of missing values.* In the analysis we have used 144 series having no missing values. Therefore, from the very beginning the selection of series was narrowed to a significant extent. For the purpose of the imputation of missing values, the expectation maximization (EM) algorithm (Stock-Watson 2002) is standardly used. This is especially useful in the so-called *nowcasting* of, for example, GDP (Schumacher and Breitung (2006) and Giannone et al. (2005)), where the data are available with a certain delay.
- *Treatment of outliers.* Considering the analysis concerns data correlation structure, series values unusually far from expectations of the (stationary) analysed series, should be excluded or somehow corrected. In this respect, series are scanned prior to analysis and outliers are identified as missing values which are then treated as described in the previous paragraph.
- *Data with different frequencies.* Sometimes it is useful to include both monthly or quarterly data in the analysis. This is done using the EM algorithm.
- *Seasonal adjustment.* Although it was not necessary, we have seasonally adjusted all series which were used for factors extraction. Furthermore, we have used only trend-cycle components of series in the analysis – therefore, the procedure should be also implemented on classic SA data. Finally, the analysis should be conducted using various SA methods.
- *Selection of series.* Series which are to be used in the analysis could be selected more rigorously, for example, on the basis of some sort of *expert knowledge* information. Furthermore, the efficiency of the use of sectoral factors (extracted out of series of hereby defined five groups of series- see appendix) in forecasting can be tested.

4 Conclusion

We have described the basic properties, goals and possibilities of methods for the compression of a large number of data in just a handful of variables. Firstly, we have defined the basics of the principal components analysis and represented it as a tool for the analysis of the covariance structure of group of variables. Secondly, we have presented a forecasting model of factor analysis which is, in the estimation phase, relies on the principal components.

Furthermore, we have checked the efficiency of these methods on Croatian data. In the first step, the (second moment) information of 144 variables relevant for Croatian economy were compressed into a handful of factors and tested for possible use in the forecasting of CPI inflation on horizons up to four quarters ahead.

Using a single factor in the forecasting exercise increases the accuracy of inflation forecasts of the benchmark model, which encourages further research into factor models.

The conducted analysis is preliminary and its results cannot be considered final. There is room for improvement of the implemented models, and key unresolved issues are explicitly enumerated and therefore the key directions for further analysis.

In this paper, factor models are tested on the problem of inflation forecasting where it has been shown that information extracted from a large number of economic variables can help in forecasting inflation. However, the methodology is described sufficiently generally and can be directly applied when forecasting other economic variables of interest which are believed to be affected by the overall state in the economy (for example, indicators of real activity or financial markets variables).

Appendix

Appendix A

Parameters estimation using principal components

Let us consider the following factor model:

$$X_{ti} = F_t \lambda_i^\tau + \varepsilon_{ti} \quad t = 1, \ldots, T, i = 1, \ldots, N, \tag{17}$$

where F_t is r- dimensional vector of factors at time t and λ_i factor loadings coefficients. In matrix notation we have:

$$X = F\Lambda^\tau + \varepsilon, \tag{18}$$

or analoguosly:

$$X^\tau = \Lambda F^\tau + \varepsilon^\tau. \tag{19}$$

Applaying the least squares method one needs to estimate F_t and the matrix Λ, given the fact that both the factors and the factor loadings are unobservable.

Given the data matrix$X$ (having $N \times T$ elements) we need to estimate $N \times r$ elements of Λ and $r \times T$ elements of F, therefore overall number of parameters to be estimated is ($N + T$)×r. In order to estimate all of the parameters it is necessary to impose certain constraints. Two scenarios are to be distinguished: $T > N$ and $T < N$. Depending on these relations we adopt one of the two following estimation strategies.

Case 1 T > N

Assuming restriction $\Lambda^\tau \Lambda = I$ we minimise the following function:

$$V(F,\Lambda) = \sum_{i=1}^{N}\sum_{t=1}^{T} \varepsilon_{it}^2 = \sum_{i=1}^{N}\sum_{t=1}^{T} (X_{it} - \lambda_i F_t^\tau)^2 = tr(X^\tau - \Lambda F^\tau)^\tau (X^\tau - \Lambda F^\tau), \tag{20}$$

where tr denotes the matrix trace operator[15]. We estimate factors from (19) applying (multivariate multiple) least squares:

$$\hat{F}^\tau = (\Lambda^\tau \Lambda)^{-1} \Lambda^\tau X^\tau = (uvjet\ na\,\Lambda) = \Lambda^\tau X^\tau, \tag{21}$$

or:

$$\hat{F} = X\Lambda. \tag{22}$$

[15] Matrix trace is defined as a sum of its diagonal elements. It can be shown that $tr(A^\tau A) = \sum_{i=1}^{N}\sum_{j=1}^{T} a_{ij}^2 = \| A \|_F^2$ is squared Frobenius norm. So it is sufficient to minimise $tr(\varepsilon^\tau \varepsilon)$.

Substituting (22) in (20) and using the properties of *tr* operator it follows:

$$\begin{aligned} V(\hat{F},\Lambda) &= \| X^{\tau} - \Lambda\hat{F}^{\tau} \|_F^2 \\ &= \| X^{\tau} - \Lambda\Lambda^{\tau}X^{\tau} \|_F^2 \\ &= tr((X - X\Lambda\Lambda^{\tau})(X^{\tau} - \Lambda\Lambda^{\tau}X^{\tau})) \\ &= tr(XX^{\tau} - X\Lambda\Lambda^{\tau}X^{\tau} - X\Lambda\Lambda^{\tau}X^{\tau} + X\Lambda(\Lambda^{\tau}\Lambda)\Lambda^{\tau}X^{\tau}) \\ &= \| X \|_F^2 - tr(\Lambda^{\tau}X^{\tau}X\Lambda). \qquad (23) \end{aligned}$$

Norm $\| X \|_F^2$ does not depend on Λ and therefore $V(\hat{F},\Lambda)$ reachs the minimum for Λ that minimises $-tr(\Lambda^{\tau}X^{\tau}X\Lambda)$ or analoguosly that maximises $tr(\Lambda^{\tau}X^{\tau}X\Lambda)$. Let us notice that this is equivalent to the problem of finding maximum variance linear combinations of components of vector X (i.e. $X\Lambda$) under $\Lambda^{\tau}\Lambda = I$. Therefore it is a principal component problem described in the second section. We mentioned that the problem is solved by taking that $\hat{\Lambda}$ having as its columns the first few eigenvectors of $X^{\tau}X$ ($N\times N$), Now (22) yields factors:

$$\hat{F} = X\hat{\Lambda}. \qquad (24)$$

Case 2 T < N

As in the first case, but using (18) and under restrictions $F^{\tau}F = I$ eigenvectors of XX^{τ} ($T\times T$) need to be calculated. Depending on the dimensions we choose between the cases.

Appendix B

List of series

Prices, exchange rates

- Crude Oil (average of Brent, WTI and Dubai Fateh)
- CPI (Total)
- Food and non-alcoholic beverages
- Alcoholic beverages and tobacco
- Clothing and footwear
- Housing, water, electricity, gas and other fuels
- Furnishing, household eq. and rouitine m. of house
- Health
- Transport
- Communications
- Recreation and culture
- Education
- Catering and accommodation services
- Miscellaneous goods and services
- Goods
- Services
- Total excl. Energy
- Total excl. Energy and Food
- Goods excl. Energy
- Food, beverages and tobacco
- Core inflation
- HWWA index (USD) – Total
- HWWA index (USD) – Total excl. energy
- HWWA_ind. Raw. Materials
- HWWA_Crude oil
- PPI
- HRK/EUR
- HRK/USD
- INEER
- IREER CPI
- IREER PPI

External

- Original dana – Exports (f.o.b.)
- Original dana – Imports (f.o.b.)
- Original dana – Balance
- Machinery and transport equipment – Exports (f.o.b.)
- Machinery and transport equipment – Imports (c.i.f.)
- Machinery and transport equipment – Balance
- Other tansportation equipment – Import (fob)
- Other tansportation equipment – Export (cif)
- Road vehicles – Exports (f.o.b.)
- Road vehicles – Imports (c.i.f.)
- Road vehicles – Balance
- PPP – Exports (f.o.b.)
- PPP – Imports (c.i.f.)
- PPP – Balance
- Oil – (Export)
- Oil – (Import)
- Export (total)
- Export (energy)
- Export (intermediate)
- Export (capital)
- Export (dur. goods)
- Export (non-dur. goods)
- Import (total)
- Import (energy)
- Import (intermediate)
- Import (capital)
- Import (dur. goods)
- Import (non-dur. goods)

Money

- Net Foreign Assets
- CBN Net Foreign Asets
  - Banks Foreign Assets
  - Banks Foreign Liabilities
- Net Domestic Assets
- Claims on central government and funds
- Claims on non banking sector
  - Claims on enterpises
  - Claims on households
- Other assets
- Broadest money
- Money supply
  - Currency outside banks
  - Deposits
    - Enterpises deposits
    - Households deposits
- Quasi money
  - Kuna deposits
    - Households kuna deposits
    - Enterprises kuna deposits
    - Fc deposits
      - Fc enterprises deposits
      - Fc household deposits
  - Bonds and money market instruments
- Broad money at fixed ER
- Claims on other domestic sectors at fixed ER
- FC deposits at fixed ER
- Credits to households
- CNB international reserves
- Net usable CNB international reserves M0
- Required reserve
  - Required reserve (in kuna)
  - Required reserve deposited with the CNB (in kuna)

Real	Labour Market
Industrial production (total)	Registered unemployment
Mining and quarrying	Newly registerd
Manufacturing	Employed from the register
Electricity, gas and water supply	Deleted from the register for reasons other than employment
Eneregy	
Intermediate goods	Total persons in employement
Capital goods	Persons in paid employment in legal entities
Durable consumer goods	Persons in paid employment in legal entities in public administration (LMN)
Non-durable consumer goods	
Total volume indicies of construction	Persons in paid employment in legal entities in industry (CDE)
Total volume indicies of construction – buildings	
Total volume indicies of construction – civil engeenering works	Persons in paid employment in craft and trades and free lances
Tourist arrivals	Insured persons – private farmers
Tourist arrivals – domestic	Active population (labour force)
Tourist arrivals – foreign	Administrative unemployment rate
Tourist nights	Nominal net wage
Tourist nights – domestic	Real net wage
Trade	Nominal gross wage
GDP	Real gross wage
Investments	Nomilan gross wage in public administration (LMN)
Consumption	Nominal gross wage in industry (CDE)
Import	Real gross wage in public administration (LMN)
Export	Real goss wage in industry (CDE)
	Nominal net wage in public administration (LMN)
	Nominal net wage in industry (CDE)
	Real net wage in public administration (LMN)
	Real net wage in industry (CDE)

LITERATURE

Angelini, E., Henry, J. and Mestre, R., 2001. "Diffusion Index-Based Infation Forecasts for the Euro Area". *European Central Bank Working Paper,* No. 61.

Artis, M. J., Banerjee, A. and Marcellino, M., 2005. "Factor forecasts for the UK". *Journal of Forecasting*, 24, 279-298.

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

Banerjee, A., Marcellino M. and Masten, I., 2006. *Forecasting Macroeconomic Variables Using Diffusion Indexes in Short Samples with Structural Change*. Mimeo.

Bernanke, B. and Boivin, J., 2003. Monetary Policy in a Data-Rich Environment. *Journal of Monetary Economics,* 50, 525-546.

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

Camacho, M. and Sancho, I., 2003. "Spanish diffusion indexes". *Spanish Economic Review*, (5), 173-203.

Camba-Méndez, G. and Kapetanios, G., 2005. "Forecasting euro area inflation using dynamic factor measures of underlying inflation". *Journal of Forecasting*, 25, 491-503.

Chicago Fed Letter, 2000. *Forecasting inflation with a lot of data The Federal Reserve Bank Of Chicago*, March 2000, Number 151.

Eickmeier, S. and Ziegler, C., 2006. "How good are dynamic factor models at forecasting output and inflation? A meta-analytic approach". *Discussion Paper Series 1: Economic Studies*, 42.

Forni, M. [et al.], 2000. "The Generalized Dynamic Factor Model: Identification and Estimation". *The Review of Economics and Statistics*, 82 (4), 540-552.

Forni, M. and Reichlin, L., 1996. "Dynamic Common Factors in Large Cross-Sections". *Empirical Economics*, 21, 27-42.

Giannone, D., Reichlin, L. and Small, D., 2005. "Nowcasting GDP and inflation: the real-time informational content of macroeconomic data releases". *Finance and Economics Discussion Series,* No. 2005-42.

Giannoni, M. and Boivin, J., 2005. "DSGE Models in a Data-Rich Environment". *Computing in Economics and Finance,* No. 431.

Grenouilleau, D., 2006. "The Stacked Leading Indicators Dynamic Factor Model: A Sensitivity Analysis of Forecast Accuracy Using Bootstrapping, European Commission Directorate-General for Economic and Financial Affairs". *Economic Papers*, No 249.

Jankov, Lj. [et al.], 2007. *The Impact of USD/EUR Exchange Rate on Inflation in CEE Countries*. Dubrovnik: The Thirteenth Dubrovnik Economic Conference.

Johnson, R. A. and Wichern, D. W., 1998. *Applied Multivariate Statistical Analysis*. New York: Prentice Hall.

Kapetanios, G. and Marcellino, M., 2003. "A Parametric Estimation Method for Dynamic Factor Models of Large Dimensions, Queen Mary". *University of London Working Paper,* No 489.

Kapetanios, G., Labhard, V. and Price, S., 2007. "Forecast combination and the Bank of England's suite of statistical forecasting models". *Bank of England working papers,* No. 323.

Marcellino, M., Stock J. H. and Watson, M. W., 2005. "A Comparison of Direct and Iterated AR Methods for Forecasting Macroeconomic Series h-Steps Ahead". *Journal of Econometrics*, 26 (7), 527-549.

Matheson, T. D., 2006. "Factor model forecasts for New Zealand". *International Journal of Central Banking*, 2, 169-237.

Schumacher, C. and Breitung, J., 2006. "Real-time forecasting of GDP based on a large factor model with monthly and quarterly data". *Bundesbank Discussion Paper, Series 1*, No. 33.

Spearman, C., 1904. "General Intelligence, Objectively Determined and Measured". *American Journal of Psychology,* 15, 201-293.

Stock, J. H. and Watson, M. W., 1989. "New Indexes of Coincident and Leading Economic Indicators". *NBER Macroeconomics Annual*, 351-393.

Stock, J. H. and Watson, M. W., 1998. "Diffusion indexes". *NBER Working Paper,* No. 6702.

Stock, J. H. and Watson, M. W., 1999. "Forecasting Inflation". *Journal of Monetary Economics,* 44, 293-335.

Stock, J. H. and Watson, M. W., 2002. "Macroeconomic Forecasting Using Diffusion Indexes". *Journal of Business and Economic Statistics,* 20, 147-162.